



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO,  
Programa de Doctorado en Ciencias Biomédicas, PDCB  
Instituto de Investigaciones Biomédicas, IIB

**Evolución del proteoma con base en la evolución del código genético,  
e incorporación evolutiva de moléculas de RNA**

**TESIS**

que para optar por el grado de  
**Doctora en Ciencias Biomédicas**

Presenta

**LIBB Miryam Palacios Pérez**

DIRECTOR DE TESIS:

**Dr. Marco Antonio José Valenzuela**

Instituto de Investigaciones Biomédicas, Grupo de Biología Teórica

COMITÉ TUTORAL:

**Dr. Arturo Carlos Il Becerra Bracho**

Facultad de Ciencias, Departamento de Biología Evolutiva

**Dr. Pedro Eduardo Miramontes Vidal**

Facultad de Ciencias, Departamento de Matemáticas, Grupo de Biomatemáticas

México  
Enero 2021



Universidad Nacional  
Autónoma de México

Dirección General de Bibliotecas de la UNAM

**Biblioteca Central**

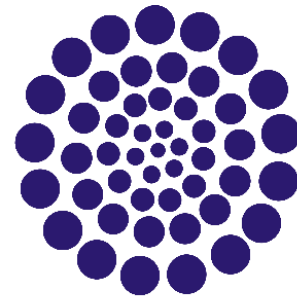


**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



**CONACYT**

DOCTORADO en  
**CIENCIAS BIOMÉDICAS**

Agradecimientos institucionales:  
Universidad Nacional Autónoma de México, UNAM  
Instituto de Investigaciones Biomédicas, IIB  
Consejo Nacional de Humanidades Ciencia y Tecnología, CONACyT

Agradecimientos Profesionales:  
Dr. Marco Antonio José Valenzuela  
Dr. Arturo Carlos Il Becerra Bracho  
Dr. Pedro Eduardo Miramontes Vidal  
CONACyT, beca de investigación **694877**  
Maestro en Ingeniería Multimedia Juan Román B.  
Dr. Francisco Prosdocimi

***GRACIAS A:***

JM Chávez Sánchez y la hermosa familia que hemos iniciado. TE AMO. Gracias!

A mis padres MLPZ e IPC, quienes han permanecido ...

Aye! Hermana y amiga.

Mi GRAN familia extendida: l@s Pérez Soria, l@s Palacios Contreras, l@s Chávez Sánchez.

Grupo de Biología Teórica.

Mis mentores y amigos: Marco, Pedro y Arturo.

Francisco P.

## ÍNDICE

1. Marco teórico .....	5
2. Estrategia experimental .....	9
3. Hallazgos generales .....	11
4. Desarrollo .....	13
4.1 Grafos fenotípicos de aminoácidos codificados por RNY (colaboración) ..	14
4.2 <i>Ur</i> -proteoma son CSBSs (autoría principal) .....	21
4.3 Proteoma antes de LUCA (autoría principal) .....	35
5. Discusión general .....	45
6. <i>EXTRA</i> , otros proyectos relacionados (moléculas de RNA) .....	53
6.1 La teoría de la información revela la evolución de los elementos de identidad del tRNA en los tres dominios de la vida (colaboración) .....	56
6.2 La historia antigua de la formación del PTC, contada por el análisis de información y la conservación (colaboración) .....	65
6.3 La evolución del RNA puede explicar la evolución de la traducción (autoría principal, aún no publicado) .....	
7. Síntesis de extra .....	82
8. Conclusiones y perspectivas .....	86
9. Referencias .....	88

## 1. MARCO TEÓRICO

La evolución de las especies mediante la selección natural puede ser representada como un proceso ramificado, que se puede rastrear hacia atrás en el tiempo (Darwin, 1859) y en cuya raíz se ubicaría el último ancestro común universal, o LUCA por sus siglas en inglés (Woese, 1998; Forterre et al., 2005). LUCA es a su vez el resultado de varios eventos evolutivos, cada uno caracterizado por la presencia de distintos componentes fundamentales: Moléculas prebióticas geoquímicas → nucleótidos (nt) + aminoácidos (aa) → ARN (o RNA) + aa y péptidos → RNA y Proteínas → RNA y Proteínas + ADN (DNA) → DNA + Proteínas + RNA (Cech, 2012), donde la última fase corresponde al contenido de LUCA.

De acuerdo con estudios de química prebiótica (Lazcano and Bada, 2003; Lazcano and Miller, 1999; Miller, 1953; Parker et al., 2014) y evidencias biogeoquímicas terrestres y de meteoritos (Miller et al., 1976), entre los componentes prebióticos principales pueden encontrarse lípidos (Janas et al., 2006; Segré et al., 2001; Sharov, 2009), cofactores nucleotídicos (Huang et al., 2000; Jadhav and Yarus, 2002; Wächtershäuser, 1988b) y cofactores metálicos (Bray et al., 2018). Dado que dichos estudios tratan de reconstruir la transición de moléculas prebióticas a proto-bióticas, se les denomina análisis abajo-hacia-arriba o "*bottom-up*".

Por otro lado, para determinar el contenido genético de LUCA, típicamente se llevan a cabo comparaciones de la mayor cantidad de genomas posibles, con el objetivo de reconstruir el conjunto original de genes comunes mediante la comparación de genomas completos de diversos organismos, teniendo en cuenta distintas variantes y escenarios, que van desde la comparación de genomas completos (Delaye et al., 2005; Gil et al., 2004; Glass et al., 2006; Itaya, 1995; Koonin, 2003; Mushegian and Koonin, 1996); así como la evolución de las superfamilias de proteínas que pudo haber contenido LUCA (Ranea et al., 2006) o los módulos mínimos de proteínas (Sobolevsky and Trifonov, 2006), hasta llegar a determinar secuencias totalmente invariantes (Goto et al., 2007). Dado que el objetivo es rastrear el contenido genético mínimo del LUCA a partir de organismos actuales, a dichos análisis comparativos se les denomina arriba-hacia-abajo o "*top-down*".

Teóricamente, se pretende que los resultados de los análisis “bottom-up” y “top-down” se encuentren en el tiempo, de manera que pueda trazarse la evolución desde el origen de la vida hasta la aparición y desarrollo del LUCA (Forterre and Gribaldo, 2007). De hecho existen algunos trabajos que originalmente se indicó que delineaban las proteínas que contenía LUCA (Sobolevsky et al., 2013), sin embargo recientes hallazgos (precisamente explicados en el presente trabajo) indican que los motivos encontrados existieron realmente antes que el LUCA.

Dada la naturaleza del LUCA, su material genético DNA plausiblemente ya seguía el código genético estándar actual (**SGC**, por sus siglas en inglés) (Koonin and Novozhilov, 2009; Woese, 1964; Woese, 2000). Pues aunque existen variantes, la gran mayoría de los organismos comparten el SGC (Watanabe and Suzuki, 2001; Watanabe and Suzuki, 2008); incluso la variante 11, que se presenta en bacterias y arqueas –organismos cercanos a la raíz del ‘árbol de la vida–, en realidad no es condición *sine qua non* de todos estos organismos (NCBI Taxonomy Homepage). Tal como se ha procurado demarcar el contenido celular de LUCA, se ha planteado trazar la evolución del código genético.

Antes de la aparición del LUCA, posiblemente existieron entidades prebiológicas denominadas *progenotes*, que habrían sido “entidades celulares muy primitivas previas a LUCA con funciones limitadas y procesos de traducción imprecisos que daban origen a proteínas estadísticas” (Woese, 1998; Woese and Fox, 1977). La molécula informacional de tales entidades biológicas era posiblemente RNA con un proto-metabolismo dirigido por éste (fenotipo=genotipo), así como cierto grado de compartimentación dictada por la membrana de tal *ribocito* (Yarus, 2002). Aquellas moléculas hereditarias probablemente eran inicialmente cortas e inestables y por ende la función más relevante de los primeros péptidos sintetizados habría sido la estabilización de moléculas ya existentes (Eigen, 1971; Woese, 1998).

Eigen y Schuster plantearon que el código genético originario o primitivo (**PGC**, por sus siglas en inglés) comprendía cadenas de ribonucleótidos que seguían un patrón **RNY** [R=purinas (A/G), Y=pirimidinas (C/U) y N=cualquiera de esos nucleótidos], funcionando como cuasi-especies “semilla” que permiten generar hiperciclos evolutivos con facilidad debido a tres características fisicoquímicas: las estructuras secundarias de RNA que se generen pueden ser reconocidas más

fácilmente por enzimas debido a que las hebras positiva y negativa son equivalentes, la simetría de ambas cadenas permite su eficiente replicación y, debido a tal simetría, ambas cadenas ofrecen funciones complementarias. El conjunto de tripletes RNY contienen tanto un codón como su reverso complementario, cumpliendo así con las características mencionadas (Eigen and Schuster, 1978).

De esta manera los primeros péptidos se conformaron por los aa: glicina (G o Gly), alanina (A o Ala), aspartato (D o Asp), valina (V o Val), serina (S o Ser), treonina (T o Thr), asparagina (N o Asn) e isoleucina (I o Ile); es decir que ya el primer código genético era redundante, porque dos codones codificaban un solo aminoácido. Es posible examinar los posibles ordenamientos algebraicos de esos 8 aa para determinar si alguno es prevalente en la evolución del PGC u observar si acaso existe una ruta a través de la cual pudo evolucionar (siguiente sección).

Matemáticamente se ha demostrado que existe un camino dicotómico desde el código RNY hasta el SGC actual. Por un lado, pudieron darse cambios en el marco de lectura debido a una replicasa rudimentaria menos precisa que la actual, que pudo iniciar la lectura en la segunda o en la tercera base, obteniendo así el conjunto RNY+NYR+YRN, denominado "Extendido 1" (Ex1); por otro lado, pudo haber transversiones en la primera o en la tercera base, obteniendo así el conjunto RNY+YNY+RNR, denominado "Extendido 2" (Ex2). Finalmente, ambas rutas de 48 codones, los códigos genéticos extendidos (**ExGCs** por sus siglas en inglés), convergen al complementarse entre sí y generar entonces el SGC *i.e.* los tripletes faltantes en un extendido son tripletes del otro extendido (José et al., 2009; José et al., 2011).

RNY, 16 codones	Extendido 1, 48	Extendido 2, 48
	RNY	RNY
	NYR	RNR
	YRN	YNY

La herramienta denominada grupo de renormalización (RG, por sus siglas en inglés) se ha utilizado para evaluar las propiedades matemáticas de un genoma. Un RG consiste en una serie de transformaciones que trasladan las propiedades estadísticas características de un sistema, a otro con diferente resolución. Si los componentes de un sistema se distribuyen siguiendo una ley de potencias (*power law*), quiere decir que durante el proceso de renormalización se verá que el



sistema presenta invariancia de escala, lo que implica que estamos ante un fenómeno autosimilar o fractal (Griffin et al., 2000; Schroeder, 1991; Wilson, 1979).

Se analizó entonces la distribución de potencias de los tripletes presentes en genomas de diversos organismos, ya sea que solamente contenían codones del tipo RNY, Ex1 o Ex2, de éstos se encontró que los genomas tienen un comportamiento fractal, *i.e* poseen invariancia de escala; ello quiere decir que sus propiedades estadísticas se fijaron desde su origen y podemos derivar uno de otro, implicando que las reglas fundamentales del código genético se han preservado a lo largo de, al menos, 3.5 mil millones de años de evolución (José et al., 2009; José et al., 2011).

Una vez establecido matemáticamente que el SGC se puede derivar de Ex1 o Ex2 y éstos a su vez del RNY, es posible mostrar la evolución del fenotipo. Siguiendo la metodología para obtener los genomas RNY, Ex1 y Ex2 de diversos organismos (José et al., 2009; José et al., 2011), se puede encontrar el proteoma –entendido como la colección completa de las proteínas de un organismo específico– correspondiente.

## 2. ESTRATEGIA EXPERIMENTAL

El código genético se puede representar como grafos, en que cada nodo corresponde a un triplete, conectado a otro nodo por un vértice que representa la mutación de un nucleótido. Para el caso del PGC de tripletes RNY, los nodos una vez conectados se substituyen con los aa correspondientes y se observa que, en todos los casos, los aa se ordenan con base en su requerimiento polar. Los grafos fueron entonces analizados con base en sus medidas de centralidad (José *et al.*, 2015, donde se colaboró con la descripción de las características fisicoquímicas de los aa estudiados).

Empero, cuando se desean analizar organismos, es necesario primero obtener un genoma que contenga solamente marcos abiertos de lectura (ORF, por las siglas en inglés de "open reading frame"); para que la estructura original de las cadenas se conservara, se concatenaron todos los ORF reportados, uno tras otro tal cual aparecen en el orden original de orientación en el cromosoma (OOO) y por ello nombrados "genomas OOO". Teniendo el genoma OOO del organismo a analizar, se utiliza un programa en Perl que únicamente preserva los tripletes que no pertenecen al conjunto determinado, en primera instancia los 16 tripletes del RNY, o los 48 del Ex1, o los 48 del Ex2 (José *et al.*, 2009). Dicho programa se puede modificar dependiendo de la ruta evolutiva a evaluar, por ejemplo, para obtener un conjunto con tripletes del tipo YNR que pueda fungir como control. Para obtener controles negativos, se llevan a cabo aleatorizaciones de los genomas OOO, deshaciendo así el sentido biológico de la secuencia. Los OOO aleatorizados se depuran también para obtener genomas con el tipo de tripletes requerido. De cada organismo deseado, se obtiene entonces su genoma OOO, éste se aleatoriza y ambos se depuran para obtener únicamente genomas formados por tripletes RNY, por Ex1 o por Ex2. Posteriormente, utilizando BLAST, se obtiene la lista de proteínas codificadas por el genoma definido y su correspondiente aleatorización (Palacios-Pérez *et al.*, 2018).

El *valor -E* es el parámetro más utilizado para filtrar los resultados de un alineamiento por BLAST, dicho parámetro describe el número de aciertos de calidad similar que se pueden encontrar al azar dada un base de datos finita (Altschul *et al.*, 1990); sin embargo, dado que el

*valor - E* depende de la longitud de la secuencia y los tripletes de cada código genético corresponden solamente a una fracción del genoma completo, únicamente fragmentos de cada proteína serán codificados por cada tipo de tripletes (RNY, Ex1 y Ex2) y por tanto no es posible establecer un valor de corte como se realizaría con proteínas más recientes para las que se preselecciona un *valor - E* restringido (tendiente a cero). Para definir entonces las proteínas que estadísticamente presentan porciones codificadas por algún tipo de triplete definido, se compara el denominado *valor - E* de los resultados de cada genoma biológico contra el *valor - E* de su control aleatorizado correspondiente. Las porciones indicadas por BLAST se ordenaron con base en la secuencia original en nt y, cuando la proteína está reconstituida, se traduce en aa (Palacios-Pérez et al., 2018). Si los genomas son varios, se reconstruyen todas las proteínas del mismo tipo a partir de sus fragmentos ordenados y las reconstrucciones se alinean para obtener consensos (Palacios-Pérez and José, 2019).

Una vez obtenidas las porciones de las proteínas codificadas por cada tipo de 'código genético' (RNY, Ex1 o Ex2), se lleva a cabo una predicción de la estructura de tales péptidos o proteínas; dependiendo de la antigüedad del código genético utilizado, la estructura podría resultar más o menos semejante a la proteína moderna correspondiente. Adicionalmente puede predecirse el tipo de moléculas que dichos péptidos o proteínas podían unir (Palacios-Pérez et al., 2018; Palacios-Pérez and José, 2019).

### 3. HALLAZGOS GENERALES

Es posible inquirir el fenotipo a lo largo de la evolución del genotipo hasta antes del LUCA, ya sea a nivel de aa o de péptidos y proteínas, *i.e.* se puede recuperar el fenotipo codificado por cada tipo de código genético, iniciando con el primordial RNY y posteriormente siguiendo la codificación mediante Ex1 y Ex2, que emergieron antes del formal surgimiento de LUCA.

Entendido el fenotipo únicamente como los rasgos observables de un individuo y no con relación al ambiente (Fenotipo | NHGRI); en este caso los aa o péptidos codificados por un tipo de código genético (el PGC o los ExGCs).

#### GRAFOS FENOTÍPICOS EN AMINOÁCIDOS, CORRESPONDIENTES AL CÓDIGO GENÉTICO RNY

Inicialmente se muestra el trabajo colaborativo acerca de los grafos fenotípicos de los 8 aa codificados por los 16 tripletes del tipo RNY. Se encontró que todos los grafos son altamente simétricos, con sólo una posible “ruptura de simetrías” en el aminoácido Ala, por lo que se deduce que el código RNY estaba *congelado*, aunque sí pudo evolucionar (José et al., 2015).

#### UR-PROTEOMA

El fenotipo correspondiente al genotipo primitivo, *i.e.* basado en el código genético RNY, se investigó utilizando el genoma de solamente un organismo —*Streptococcus agalactiae* A909, bacteria que se inicialmente se utilizó para establecer el concepto de pan-genoma, que es el conjunto completo de todos los genes de las diferentes cepas que pertenecen a una misma especie (Medini et al., 2005; Tettelin et al., 2005)—, a pesar de lo cual fue posible revelar que las proteínas modernas que contienen regiones codificadas por tripletes del tipo RNY pertenecen a diversas rutas metabólicas y cumplen distintas funciones celulares; sin embargo, contrario a lo que se considera y lo esperado, dichos módulos proteicos fundamentales no constituyen los sitios catalíticos de las proteínas modernas, sino que corresponden a sitios de unión y estabilización de cofactores —**CSBSs** por las siglas en inglés de **Cofactor Stabilising Binding Sites**— de manera que podemos decir que los primordios de las proteínas actuales, el **Ur-proteoma**, constituye un

***bindoma*** (por la palabra inglesa "binding" que hace referencia a unión) para moléculas que se han propuesto entre las primeras moléculas prebióticas (Palacios-Pérez et al., 2018).

Con la metodología descrita fue posible además reconstruir la estructura de ciertos "módulos de proteínas" que se indicaron como los más antiguos presentes en el LUCA (Sobolevsky and Trifonov, 2006); sin embargo, por sus características, se observa que son en realidad CSBSs, que por tanto existieron mucho antes que el LUCA y son capaces de unir moléculas protobióticas.

#### EVOLUCIÓN DEL PROTEOMA, CODIFICADO POR CÓDIGOS EXTENDIDOS

Es posible trazar la evolución de las proteínas que fueron constituyendo el repertorio del LUCA, observando el fenotipo codificado por cada uno de los genotipos que aparecieron a lo largo de la evolución; de esta manera, se inicia con el Ur-proteoma codificado por el PGC, de tripletes RNY, para continuar con los proteomas codificados por cada uno de los ExGCs (Ex1 y Ex2).

Utilizando los genomas de 26 especies (13 bacterias y 13 arqueas) de diversos estilos de vida y características genómicas distintas (mesófilos y termófilos, piezotolerantes y radiotolerantes, parásitos y de vida libre, con genoma lineal y con genoma circular, con genomas muy grandes y con genomas muy pequeños, etcétera), fue posible trazar la evolución proteica gradual; así como predecir la estructura y los posibles ligandos de las proteínas recuperadas pertenecientes a dichos *proteomas extendidos* (Palacios-Pérez and José, 2019).

De manera relevante, se encontró que todas las proteínas del proceso de traducción estaban ya presentes antes de LUCA; adicionalmente, se encontraron porciones importantes de proteínas que actualmente forman parte de organismos extremófilos y no solamente se infiere su existencia como anteriormente se había hecho (Weiss et al., 2016).

## 4. DESARROLLO

### 4.1 Grafos fenotípicos de aa codificados por RNY (colaboración)

“The 12 different types of graphs of the 8 amino acids encoded by the presumably primeval RNY code are derived. The symmetry groups of these graphs are analyzed and coincide with the corresponding values of polar requirement for each amino acid. The symmetry groups at the codon level are partially carried over as a group or subgroup at the amino acid level. Measures of centrality of the 12 graphs indicate that all amino acids were equally relevant irrespective of its chronological order of its appearance. The elimination of any amino acid would be strongly selected against and therefore the genetic code at this stage was already frozen” (José et al., 2015).

### 4.2 *Ur*-proteoma son CSBSs (autoría principal)

“Herein we outline a plausible proteome, encoded by assuming a primeval RNY genetic code. We unveil the primeval phenotype by using only the RNA genotype; it means that we recovered the most ancestral proteome, mostly made of the 8 amino acids encoded by RNY triplets. By looking at those fragments, it is noticeable that they are positioned, not at catalytic sites, but in the cofactor binding sites. It implies that the stabilization of a molecule appeared long before its catalytic activity, and therefore the *Ur*-proteome comprised a set of proteins modules that corresponded to Cofactor Stabilizing Binding Sites (CSBSs), which we call the primitive bindome. With our method, we reconstructed the structures of the "first protein modules" that Sobolevsky and Trifonov (2006) found by using only RMSD. We also examine the probable cofactors that bound to them. We discuss the notion of CSBSs as the first proteins modules in progenotes in the context of several proposals about the primitive forms of life” (Palacios-Pérez et al., 2018).

### 4.3 Proteoma antes de LUCA (autoría principal)

“The attempt to delineate the essential features that characterized life in its beginnings and the understanding of how those features evolved, represent important scientific challenges. While there have been varied efforts in the elucidation of how the first biomolecules arose from a prebiotic environment, there remains important gaps towards the characterization of the complete repertoire of the Last Universal Common Ancestor (LUCA). We portray a step-wise proteome evolution, by looking at the phenotype encoded by each one of the genetic codes that were appearing along evolution, beginning with the primeval genetic code and then with two intermediate Extended RNA codes, which finally shaped the current Standard Genetic Code (SGC). Notably, all molecules involved in translation, such as ribosomal proteins and all tRNA synthetases, were already present before LUCA. The metabolism belonged to extremophiles as hinted by the presence of reverse gyrase and acetyl coenzyme A synthase. Furthermore, we predict the structure and possible ligands of the proteins retrieved. We have forged a bridge between the hitherto unknown proteome of progenotes and the proteome of LUCA” (Palacios-Pérez and José, 2019).

# Symmetrical and Thermodynamic Properties of Phenotypic Graphs of Amino Acids Encoded by the Primeval RNY Code

Marco V. José · Gabriel S. Zamudio ·  
Miryam Palacios-Pérez · Juan R. Bobadilla ·  
Sávio Torres de Fariás

Received: 5 October 2014 / Accepted: 26 January 2015  
© Springer Science+Business Media Dordrecht 2015

**Abstract** The 12 different types of graphs of the 8 amino acids encoded by the presumably primeval RNY code are derived. The symmetry groups of these graphs are analyzed and coincide with the corresponding values of polar requirement for each amino acid. The symmetry groups at the codon level are partially carried over as a group or subgroup at the amino acid level. Measures of centrality of the 12 graphs indicate that all amino acids were equally relevant irrespective of its chronological order of its appearance. The elimination of any amino acid would be strongly selected against and therefore the genetic code at this stage was already frozen.

**Keywords** Phenotypic graphs · Primeval RNY code · Amino acids · Polar requirement · Symmetry groups

## Introduction

Among the organic molecules that are constituents of cells, amino acids play a prominent role as building blocks of proteins which are the quintessence of the phenotypic expression. There has been clear evidence for prebiotic formation of amino acids since the experiments of Miller (Miller 1953, 1957; Miller and Orgel 1974), which involved electrical discharges in a mixture of atmospheric gases and have been shown to produce 10 of the amino acids used in modern proteins, G – Gly, A – Ala, V – Val, D – Asp, E – Glu, P – Pro, S – Ser, L – Leu, T – Thr, I – Ile plus many other organic molecules. Seven out of these 10 amino acids are encoded by the

---

Paper presented at ORIGINS 2014, Nara Japan, July 6–11 2014.

M. V. José (✉) · G. S. Zamudio · M. Palacios-Pérez · J. R. Bobadilla  
Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, México, D.F. 04510, Mexico  
e-mail: marcojose@biomedicas.unam.mx

S. T. de Fariás  
Centro de Ciências Exatas y Naturales, Universidade Federal da Paraíba, João Pessoa, Brazil

primeval genetic code composed by RNY (R-purine, Y- pyrimidine, N-any nucleotide) codons proposed by Eigen almost 30 years ago (Eigen and Schuster 1977). This slightly degenerate RNY code comprises the following amino acids: G, A, V, D, S, T, I, N – Asn. The origin and evolution of the Standard Genetic Code (SGC) has been examined by using group theory which is a branch of mathematics to determine the symmetries of an object (Coxeter 1973). In particular, it has been shown that the putative primeval RNY code can be represented in a highly symmetrical four-dimensional hypercube (José et al. 2007). It has also been shown that by frame-shift reading mistranslations and/or by transversions in the first or third nucleotide of the RNY codons the 4-dimensional hypercube replicates, together with the appearance of new amino acids, until it generates the whole SGC (José et al. 2007, 2009).

In regard to symmetries the RNY code and the whole SGC display a primitive algebraic structure known as the Four-Klein Group, which is the only non-cyclic group. More recently, it has also been shown that depending upon the ordering of the 4 nucleotides A, U, G, and C, there are 24 ways to represent the SGC (José et al. 2012) and there are only 12 ways to represent each of the corresponding phenotypic graphs of amino acids (network of amino acids as encoded by codons taking into account the structure of the genetic code (José et al. 2014)). All graphs exhibit disjoint clusters of amino acids when their polar requirement values are used (José et al. 2014). The polar requirement is a measure of chemical properties centering on the chromatographic mobility of amino acids (Woese et al. 1966). The genetic code seems to be organized so that common substitutions cause little changes in this property (Freeland and Hurst 1998).

In this work we pose the following questions: Given that the RNY code exhibit certain symmetries, do they carry over to their corresponding phenotypic graphs of the 8 primeval amino acids? Do the polar requirement values still form clusters in the networks of amino acids?

Herein, we briefly describe the main physicochemical properties of the 8 primeval amino acids. Second, we provide some algebraic definitions for understanding the type of symmetries of a graph. Third, we present the different types of primeval graphs of amino acids together with their corresponding polar requirement values. Next, we analyze the graph topologies, the symmetry groups of each of them, and we calculate the centrality measures of the 12 graphs. Finally, we briefly discuss the present findings emphasizing its value for supporting an RNY code and its phenotypic networks of amino acids (José et al. 2014).

## Physicochemical Properties of the Primeval Amino Acids

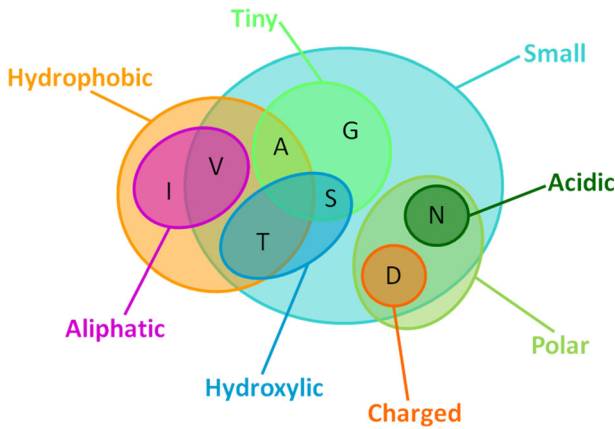
Note in Fig. 1, that all amino acids are, depending on the length of its side chain, either small (G, A, V, D, S, T, N) or even tiny (A, G, S) except I. There are 4 hydrophobic (A and T, and I and V (both aliphatic)), 2 hydroxylic (S and T), and 2 polar (N acidic and D charged). Therefore, in this small set of amino acids one can already find a wide repertoire of physicochemical properties. As we will see in the section on symmetry groups, the property of polar requirement exhibits a symmetrical pattern together with the symmetries of the graphs of these 8 amino acids.

## Definitions

### *Graph Automorphism*

A graph automorphism is a function  $f: V(G) \rightarrow V(G)$  that is bijective, i.e., it has a one-to-one mapping, and preserves edge-vertex connectivity, this means that for  $a, b \in V(G)$  if  $a$  and  $b$  are joined by an edge then  $f(a)$  and  $f(b)$  are also joined by an edge.





**Fig. 1** Venn diagram showing the physicochemical properties of amino acids encoded by RNY codons

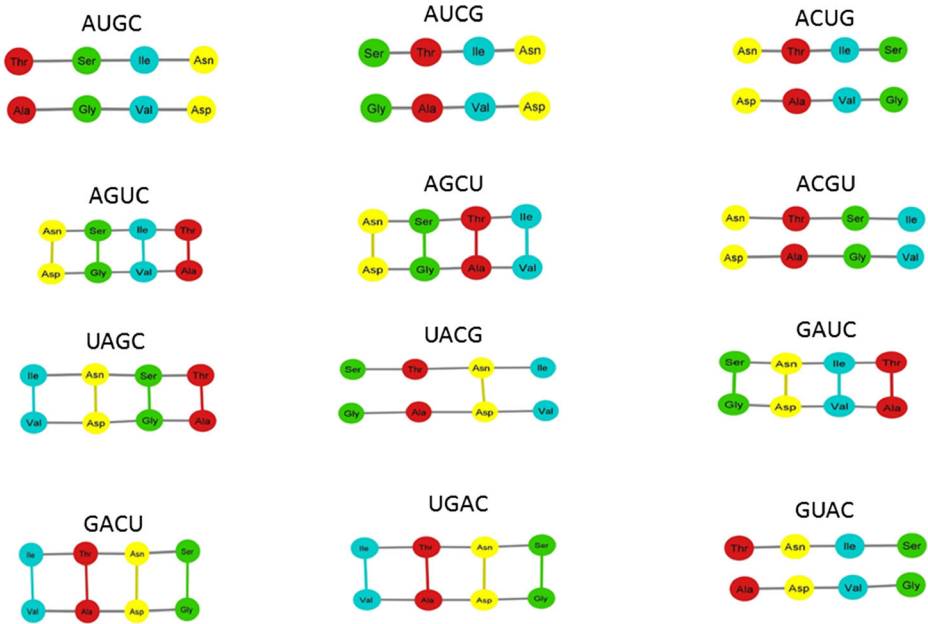
Given the set  $N = \{A, U, G, C\}$  of the 4 RNA nucleotides, the set of all the triplets is the set  $NNN = N^3 = \{xyz | x, y, z \in N\}$ , that comprises the 64 codons of the SGC. Let  $A$  be the set of all the amino acids encoded by the RNY code which is a subset of  $N^3$ , here dubbed  $N'$ . The RNY codons are the vertices of a graph  $K$ , in which the edges are given by the algebraic structure of the set  $N^3$ .

Now we define in the set  $N'$  the following equivalence relation: for  $x, y \in N'$   $x \mathcal{R} y \Leftrightarrow x$  and  $y$  encode the same amino acid in  $A$ . This set is now partitioned according to the amino acid that each codon encodes, and now it is straightforward to make the quotient  $N'/\mathcal{R}$ . This algebraic operation converts the set  $N'$  in a graph  $G$  where the set of vertices  $V(G)$  will be the set of amino acids present in the RNY code and the set of edges  $E(G)$  will be defined in the following manner: two vertices will be joined by an edge if there exists two codons in the graph  $K$ , that encode those amino acids that are also joined by an edge. The graph  $G$  is constructed according to the topology of the set  $N^3$  and since it can be constructed in 24 ways which are given by the permutations of the set  $N$ , but only 12 of those permutations yield to different graphs  $G$ , then we have 12 graphs  $G_i$  with  $i \in \{1, 2, 3, \dots, 12\}$ .

## Results

### Graph Topology

When the polar requirement scale values of each amino acid are considered, the vertices of each graph  $G_i$  (see Fig. 2) are colored according to this scale and it turns out that amino acids are neatly separated into 4 major groups with 2 amino acids of the RNY code in each of these groups. Note also in Fig. 2 that the 12 graphs  $G_i$  display the same core structure: two strands of four vertices each but the colors of the vertices in each strand are exactly the same. Hence we have a situation in which one amino acid of each color is present in each strand and we have the same set of amino acids on each strand regardless of the chosen graph. The strands are  $S_1 = \{\text{Thr, Ser, Ile, Asn}\}$ ,  $S_2 = \{\text{Ala, Gly, Val, Asp}\}$ , where Thr and Ala are of the red group, Ser and Gly are of the green group, Ile and Val are of the blue group, and Asn and Asp are of the yellow group. Another characteristic is that for each group of polar requirement the amino



**Fig. 2** The twelve graphs of the amino acids encoded by RNY codons. Amino acids are colored according to its polar requirement score

acids with higher scales conform the strand  $S_1$ , and the ones with the lowest scales conform the strand  $S_2$ .

A part of the topology of a graph is given by its connectedness. If the graph is connected then it is possible to move from a fixed vertex to any other vertex following a sequence of adjacent edges. In the graphs  $G_i$ , 5 (AUGC, AUCG, ACUG, ACGU, and UACG) out of the 12 are connected while the remaining 7 (AGUC, AGCU, UAGC, GAUC, GACU, and UGAC) are disconnected, and in the disconnected ones the connected components are exactly the strands  $S_1$  and  $S_2$ .

### Symmetries Groups of the 12 Graphs

The set of automorphisms of a graph does have a group structure. In Table 1, a list of the different symmetry groups of the 12 graphs of amino acids is shown: for 6 (AGUC, AGCU, UAGC, GAUC, GACU and UGAC) out of the 12 graphs  $G_i$ , the group is the Klein Four-Group, for 5 (AUGC, AUCG, ACUG, ACGU, and GUAC) out of the 12, the group is the so-called Dihedral 4 group denoted by  $Dih_4$ , and for only 1 (UACG) graph the group is the binary set  $\mathbb{Z}_2 = \{0, 1\}$ . The Klein Four-Group arises on the counting of the symmetries of the rectangle which consists of two perpendicular reflections, the group  $Dih_4$  describes the complete symmetries of a square which comprises a rotation and a reflection through one diagonal of the square, and the group  $\mathbb{Z}_2$  possesses only one reflection. It is noteworthy to mention that symmetries at the codon level of the genetic code have been analyzed and the Klein Four-Group is the one that reflects these symmetries and the same group arises to describe the symmetries of graphs of the primeval amino acid since this group is also a subgroup of  $Dih_4$ .

**Table 1** General topological properties of each graph

Ordering	Connected	Group automorphism
AUGC	No	Dih <sub>4</sub>
AUCG	No	Dih <sub>4</sub>
ACUG	No	Dih <sub>4</sub>
AGUC	Yes	Klein Four-Group
AGCU	Yes	Klein Four-Group
ACGU	No	Dih <sub>4</sub>
UAGC	Yes	Klein Four-Group
UACG	Yes	Z <sub>2</sub>
GAUC	Yes	Klein Four-Group
GACU	Yes	Klein Four-Group
UGAC	Yes	Klein Four-Group
GUAC	No	Dih <sub>4</sub>

Graph Centrality Measures

The statistical centrality measures of the 12 graphs  $G_i$  are shown in Table 2. The corresponding average of these estimates yielded the same value for the 2 amino acids in each polar requirement group, except for closeness in the red group composed by *Ala* and *Thr* which reflected different values. The degree is practically equal to 2 for the 8 amino acids, which means that all amino acids are uniformly connected to each other. Since all amino acids have the same value of the eigenvector this implies that all amino acids are equally relevant. Betweenness is the number of shortest paths between pairs of vertices that pass through a given vertex. In this case, betweenness measures how many times an amino acid lies on the shortest path across all the other amino acids in a graph. Closeness reflects the average distance of a vertex to all others. It can be envisaged as how long it would take to spread information from a given amino acid to all the other amino acids. Consequences of mutations and other errors in transmitting genetic information are ameliorated not only by the structure of the RNY code but also by the symmetries of the network of amino acids. *Ala* stands out as the amino acid with the smallest closeness centrality to the remaining amino acids, which has the largest path distances to the other amino acids.

**Table 2** Average centrality measures

Amino acid	Polar requirement group	Degree	Closeness	Eigenvector	Betweenness
Serine	Green	2	0.687	0.335	4.972
Glycine	Green	2	0.687	0.335	4.972
Threonine	Red	2	0.678	0.336	4.694
Alanine	Red	2	0.056	0.336	4.694
Asparagine	Yellow	2.083	0.71	0.353	6.972
Aspartic Acid	Yellow	2.083	0.71	0.353	6.972
Isoleucine	Blue	2	0.686	0.343	4.027
Valine	Blue	2	0.686	0.343	4.027

## Conclusions

The phenotypic graphs of the RNY code possess high structural symmetries which are given by the group of automorphisms present on each arrangement of the graphs  $G_i$ . The symmetry group at the codon level partially carries over as a group or subgroup at the amino acid level. The  $Dih_4$  and  $\mathbb{Z}_2$  groups are new elements of symmetry in the phenotypic graphs. This is also reflected in the amino acids centrality measures which cluster each group by the scales of polar requirement regardless of the chosen measure and the type of graph. We do rarely obtain such symmetries and centrality estimates from the structure of a biological graph. However, the centrality of *Ala* is an outlier which will facilitate further symmetry breakings. The relevance of the 8 amino acids is the same according to its eigenvalue and given a constant degree of 2 all amino acids are evenly connected among them. According to the betweenness of the graphs any chosen amino acid can be of least length across all the remaining ones. It should not be a surprise that *Ala*, *Gly*, *Val*, and *Asp*, happen to be the most abundant amino acids formed in Miller's experiments or found in meteorites (Miller 1953; Parker et al. 2011; Bada 2013).

The biological implications of these results are the following: At this stage of evolution of the genetic code, all amino acids were equally influential irrespective of the precise chronology of its appearance. The primeval RNY code was already frozen. All amino acids were equally relevant across all graphs except one in which the yellow group acted as a bridge between the two strands. Further evolution could only be achieved by symmetry breakings (José et al. 2007, 2009), which allowed the incorporation of new players into the graphs of amino acids. Graphs of the currently encoded 20 amino acids still show vestiges of symmetries like the ones found in this work (José et al. 2014).

**Acknowledgments** MVJ was financially supported by PAPIIT-IN 224015, UNAM, México.

## References

- Bada JL (2013) New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. *Chem Soc Rev* 42(5):2186–2196. doi:10.1039/c3cs35433d
- Coxeter HSM (1973) *Regular polytopes*, 3rd edn. Dover Publication Inc., New York
- Eigen M, Schuster P (1977) The hypercycle. A principle of natural organization. Part A: Emergence of the hypercycle. *Naturwissenschaften* 64:541–565
- Freeland SJ, Hurst LD (1998) The genetic code is one in a million. *J Mol Evol* 47:238–248
- José MV, Morgado ER, Govzensky T (2007) An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull Math Biol* 69:215–243
- José MV, Goveznsky T, García JA, Bobadilla JR (2009) On the evolution of the standard genetic code: vestiges of critical scale invariance from the RNA world in current prokaryote genomes. *PLoS One* 4. doi:10.1371/journal.pone.0004340
- José MV, Morgado ER, Govzensky T (2011) Genetic hotels for the standard genetic code: evolutionary analysis based upon novel three-dimensional algebraic models. *Bull Math Biol* 73:1443–1476
- José MV, Morgado ER, Guimarães RC, Zamudio GS, de Farias ST, Bobadilla JR, Sosa D (2014) Three-dimensional algebraic models of the tRNA code and the 12 graphs for representing the amino acids. *Life* 4:341–373
- Miller SL (1953) Production of amino acids under possible primitive Earth conditions. *Science* 117:528–529
- Miller SL (1957) The mechanism of synthesis of amino acids by electric discharges. *Biochim Biophys Acta* 23: 480–489

Miller SL, Orgel LE (1974) *The origins of life on the earth*. Prentice-Hall, New Jersey

Parker ET, Cleaves HJ, Dworkin JP, Glavin DP, Callahan M, Aubrey A, Lazcano A, Bada JL (2011) Primordial synthesis of amines and amino acids in a 1958 Miller H<sub>2</sub>S-rich spark discharge experiment. *Proc Natl Acad Sci U S A* 108(14):5526–5531. doi:[10.1073/pnas.1019191108](https://doi.org/10.1073/pnas.1019191108)

Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. *Proc Natl Acad Sci U S A* 55:966–974

## A Proposal of the *Ur-proteome*

Miryam Palacios-Pérez<sup>1</sup> · Fernando Andrade-Díaz<sup>1</sup> ·  
Marco V. José<sup>1</sup>

Received: 10 August 2017 / Accepted: 24 October 2017 /

Published online: 10 November 2017

© Springer Science+Business Media B.V. 2017

**Abstract** Herein we outline a plausible proteome, encoded by assuming a primeval RNY genetic code. We unveil the primeval phenotype by using only the RNA genotype; it means that we recovered the most ancestral proteome, mostly made of the 8 amino acids encoded by RNY triplets. By looking at those fragments, it is noticeable that they are positioned, not at catalytic sites, but in the cofactor binding sites. It implies that the stabilization of a molecule appeared long before its catalytic activity, and therefore the *Ur-proteome* comprised a set of proteins modules that corresponded to *Cofactor Stabilizing Binding Sites (CSBSs)*, which we call the primitive *bindome*. With our method, we reconstructed the structures of the “first protein modules” that Sobolevsky and Trifonov (2006) found by using only RMSD. We also examine the probable cofactors that bound to them. We discuss the notion of CSBSs as the first proteins modules in progenotes in the context of several proposals about the primitive forms of life.

**Keywords** Ur-proteome · Last universal common ancestor · Primeval RNY code · Cofactor stabilizing binding sites · Ligand binding

### Introduction

Since the famous publication of Charles Darwin, *On the Origin of Species by Means of Natural Selection* in 1859, the notion of common descent has sparked discussions about the nature and even the existence of a Last Universal Common Ancestor or LUCA (Doolittle and Brown 1994; Woese 1998; Penny and Poole 1999; Hoenigsberg 2003; Forterre et al. 2005;

---

Paper presented at the International Conference on the Origin of Life, San Diego, 2017.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11084-017-9553-2>) contains supplementary material, which is available to authorized users.

---

✉ Marco V. José  
marcojose@biomedicas.unam.mx

<sup>1</sup> Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, C.P. 04510 Ciudad de México CDMX, Mexico

Delaye et al. 2005; Poole and Logan 2005; Wong et al. 2007; Becerra et al. 2007; Mushegian 2008; Glansdorff et al. 2008, 2009; Kim and Caetano-Anollés 2011; Morange 2011; Goldman et al. 2013; Di Mauro et al. 2014). The concept of LUCA is of paramount importance, and it is conceived as a virtual entity that possessed a DNA-based genome as well as complex processing information pathways already developed (Forterre and Philippe 1999; Di Giulio 2003). Before LUCA, the RNA was the informational molecule, its proto-metabolism was directed by itself, and the compartmentalization by a ribocyte membrane was already present (Yarus 2002). Woese (1998) developed the concept of “progenotes” i.e. primitive cells before LUCA, with limited functions and an imprecise translation process, which eventually originated LUCA.

There have been several proposals towards the elucidation of the origin and evolution of every biological trait, from prebiotic molecules to the emergence of the life as we know it. According to (Cech et al. 2006, Cech 2012), the most accepted chronology of the evolutionary events, that occurred well before the emergence of LUCA, is: prebiotic molecules → RNA world, when ribozymes performed all catalytic activities → ribonucleoprotein (RNP) world, that includes the peptides generated by means of RNA → LUCA, that we could refer as DNA world based on the current standard genetic code (SGC).

Eigen and Schuster (1978) proposed a primeval genetic code based on ribonucleotides of repeating RNY triplets (where R means purine, Y means pyrimidine, and N any of them). Accordingly, they would encode peptides made of the amino acids (aa): glycine (Gly or G), alanine (Ala or A), aspartate (Asp or D), valine (Val or V), serine (Ser or S), threonine (Thr or T), asparagine (Asn or N) and isoleucine (Ile or I). Lehman (2002) arrived at the same conclusions about RNY as the first genetic code using a stereo-chemical model. Yang (2005) also proposed a primeval genetic code based on the stereo-electronic properties of nucleotides (nt) and aa, but his model includes proline as an ancient aa.

The SGC has been theoretically derived from the primeval RNY genetic code under a model of sequential symmetry breakings (José et al. 2007, 2011), and vestiges of this primeval genetic code were found in current genomes of both Eubacteria and Archaea (José et al. 2009). All distance series of codons showed critical scale invariance not only in RNY sequences (all ORFs concatenated discarding the non-RNY triplets), but also in codons of two intermediate steps, called extended I or extended II, of the genetic code and of codons in current genomes (José et al. 2009, 2011). An astounding result is that such scale invariance has been preserved for at least 3.5 billion years, beginning with an RNY genetic code to the SGC throughout the two intermediate steps.

Based on: 1. The progenote community proposed by Woese (1998) 2. The proposal by Eigen and Schuster (1978) about the primitive translation system 3. The algebraic derivation of SGC from RNY code (José et al. 2007, 2009, 2011) 4. The description of the thermodynamic properties of the phenotypic graphs of the aa encoded by the RNY model (José et al. 2015) and 5. The concept that the conservation degree of a protein sequence is virtually proportional to its antiqueness (Pagel and Pomiankowski 2007), we searched for vestiges of the ancient phenotype (made of aa GADVSTNI), encoded in the primeval genetic code.

We began by extracting an early proteome (i.e. only the open reading frames, ORFs) from a contemporary organism, from which we recovered the fragments of the proteins coded by RNY triplets by sequence alignment. Once revisited the information about the specific proteins that were retrieved, we found that sequences are not located in catalytic sites but in cofactor binding sites. Next, using different algorithms to predict structures and the ligands they bind, we revealed that the first peptides were **cofactor stabilising binding sites (CSBSs)** and

constitute a collection that we called the *bindome*. This bindome coincides with the findings of Sobolevsky and Trifonov (2006) about the “first protein modules”. Additionally, the kind of peptides encoded by RNY agrees with the ones proposed by the Self-Referential Model (SRM) (Guimarães et al. 2008). Our results are also consistent with the notion that large informational molecules probably did not have sufficient stability at the early stages of life, so that the synthesized proteins, of no more than 25 to 30 aa length, performed a crucial role in the binding and stabilization rather than achieving complex enzymatic functions, i.e. binding-domains that can still be found in modern proteins (Eigen and Schuster 1978; Woese 1998; Trifonov 2006).

## Methods

First, we obtained the .ffn sequences from the bacterial genome *Streptococcus agalactiae* strain A909 (*SagA909*). To conserve the original structure of the genome, all ORFs were assembled in the original order and orientation (OOO, as described by José et al. 2009). As a negative control, a random sequence from the OOO genome was generated by shuffling the nucleotides, to eliminate the biological sense of the sequence. We developed a *script* to discard all non-RNY triplets from the OOO sequence and from its corresponding shuffled control. The resulting RNY sequence was compared with the original genomic sequence with an ad hoc computational algorithm that calls BLAST programme (Altschul et al. 1990). For the output file, the parameters were not restricted to any threshold (E-value, sequence length or identity percentage). Instead, to establish a cut-off value for the obtained proteins, a numerical comparison of the parameters of the biological sense sequence and of random sequence was made. We discarded the proteins that resulted from the E-value comparison between biological and random sequences. The remaining proteins were ranked according to their functional category, combining the information from COG (Tatusov et al. 2003) and KEGG (Kanehisa et al. 2006) databases, as well as the links in Protein Clusters collection (<http://www.ncbi.nlm.nih.gov/proteinclusters>), and pertinent literature about each identified protein. For each protein, the portions coded by RNY triplets was indicated, and they were translated with the EMBOSS programme “sixpack” ([http://www.ebi.ac.uk/Tools/st/emboss\\_sixpack/](http://www.ebi.ac.uk/Tools/st/emboss_sixpack/)), and the actual string of aa for each protein was recognised. To establish a negative control for each fragment, the whole sequence of aa of the corresponding protein was shuffled thrice and a fragment of the same length as the biological one was obtained.

To verify the function of the fragments recovered by the BLAST procedure, we sequentially introduced them into the pertinent on-line programmes of the suite Zhang Lab (<https://zhanglab.ccmb.med.umich.edu/>). Given the requirements of those programmes, if the matched fragment has less than 10 aa, some context aa from the modern protein must be added to fill the missing ones. The first approximation was made with I-TASSER programme (Zhang 2008), considered as a leader suite in ab initio structural protein predictions; the best output model, indicated by its C-score (the best is the highest in the range 0 to 1), was relaxed with ModRefiner programme (Xu and Zhang 2011). Finally, the structures and their probable ligands were obtained with the meta-server COACH (Yang et al. 2013). Only the structures whose probable ligands existed in Hadean times were retrieved (even if their C-score is not as high as can be expected for a typical modern protein), because on the early origins of life the complex molecules were not stable, according to several proposals (Miller 1953, Wächtershäuser 1988, Yarus 2002, 2011, Szathmáry 2007). Finally, the structures were



visualised with Chimera (Pettersen et al. 2004). The same algorithm was done for the random fragments, establishing a control for each protein.

## Verification of Sequences and Structures

We reconstructed the structures of the “protein modules conserved since LUCA” as reported by Sobolevsky and Trifonov (2006), and we also predicted the most likely ligands that those modules could bind. Some of their octameric sequences were considered, and we reconstructed their structures using the same algorithm that we used with our own sequences. In some cases, it was necessary to add one Gly at each terminus to fulfil the requirement of a minimum of 10 aa for the *ab initio* programme. The structures without ligands were first depicted as “traces” using the VMD programme (Humphrey et al. 1996), to facilitate the visual comparison between our reconstructions and the structures found by Sobolevsky and Trifonov (2006) (the portion highly conserved is marked with red in that paper). Thereafter, we obtained the probable ligands for each module, visualising the results with Chimera (Pettersen et al. 2004).

## Results

### Sequence Alignment Results

The E-value considers the length both sequences, and given that an RNY sequence is approximately one third of the total length, it is not convenient to restrain a priori the parameters. Therefore, after a numerical comparison of the relative frequency of their E-values, the cut-off was set at  $E \leq 0.05$ . More than one hundred proteins encoded in *SagA909* genome contain a region codified by RNY triplets. Those proteins were categorised as follows: RNA processing, DNA processing, Metabolism, Proteins processing, Membrane or wall metabolism, Resistance and Survival, Transporters and Channels, Transduction, Hypothetical and/or conserved proteins.

From the randomised genome only one putative protein was captured within the threshold. Beginning with sequence alignments, the proteins recorded are highly heterogeneous and they apparently cover the whole spectrum of current functionalities. However, by analysing those fragments codified by RNY, it is noteworthy that those segments are situated at cofactor binding sites rather than in catalytic sites. Among all the proteins that contain a fragment coded by RNY triplets we have, for instance: the ATP-synthase alpha-subunit (*ATP- $\alpha$* ), which does not perform any catalytic reaction, but it presumably corresponds to the “ATP binding site”; the region that matched with the ribosomal protein L6 (*rL6*), possibly interacts with RNA; the region that matched with the DNA topoisomerase (*DNA $topo$* ) possibly binds metallic cations or nt; it is surprising the finding of a protein that belongs to the family metallo-beta-lactamases (*M $\beta$ L*), however the region that matched probably binds metal cations; one of the hypothetical proteins (*Hyp1*), that contain a region coded by RNY triplets, belongs to the superfamily transpeptidase, but the region that matched is not close to the active site. For a selected list of identified proteins that acted as CSBS see Supplementary Information I.

Given that the identified protein from the random genome is only putative, we could not establish which region could have the binding site.

Thus, it seems that the prebiotic aa interacted with RNA chains made of concatenated RNY triplets, from which the very first ancestral peptides were encoded, generating thus a phenotype

of 8 aa, implying the appearance of the first primeval genetic code. Notice that some motifs in certain proteins are not ancient, but they can be identified because they contain aa encoded by RNY triplets.

### Structural Prediction Results for RNY Rich Fragments

Besides the sequence information, the structure of peptides or proteins can reveal more precisely their function. Therefore, after the structural reconstruction using sequentially the Zhang Lab suite, we corroborated that peptides coded by RNY triplets can bind prebiotic cofactors.

We remark that the structural prediction programs of the Zhang Lab suite yield results based on the C-score which is less error-prone in local structural alignments than the traditional RMSD. RMSD works better in global structural alignments and with proteins which are approximately of the same length, although I-TASSER gives the C-score value as well as the corresponding estimated RMSD (Zhang 2008).

Given that our RNY-coded fragments are smaller than the whole protein, it is befittingly the use of the C-score value, particularly when comparisons between biological fragments with its corresponding random sequences are made. We emphasize that the C-scores for the initial structural predictions for biological sequences are significantly different from its corresponding randomisations; for instance: for the *ATP- $\alpha$* ,  $C = -0.68$  (RMSD =  $2.0 \pm 1.6$ ), and for its randomized control  $C = -1.22$  (RMSD =  $3.0 \pm 2.1$ ); for *M $\beta$ L*,  $C = -0.51$  and for its randomized control  $C = -0.62$ .

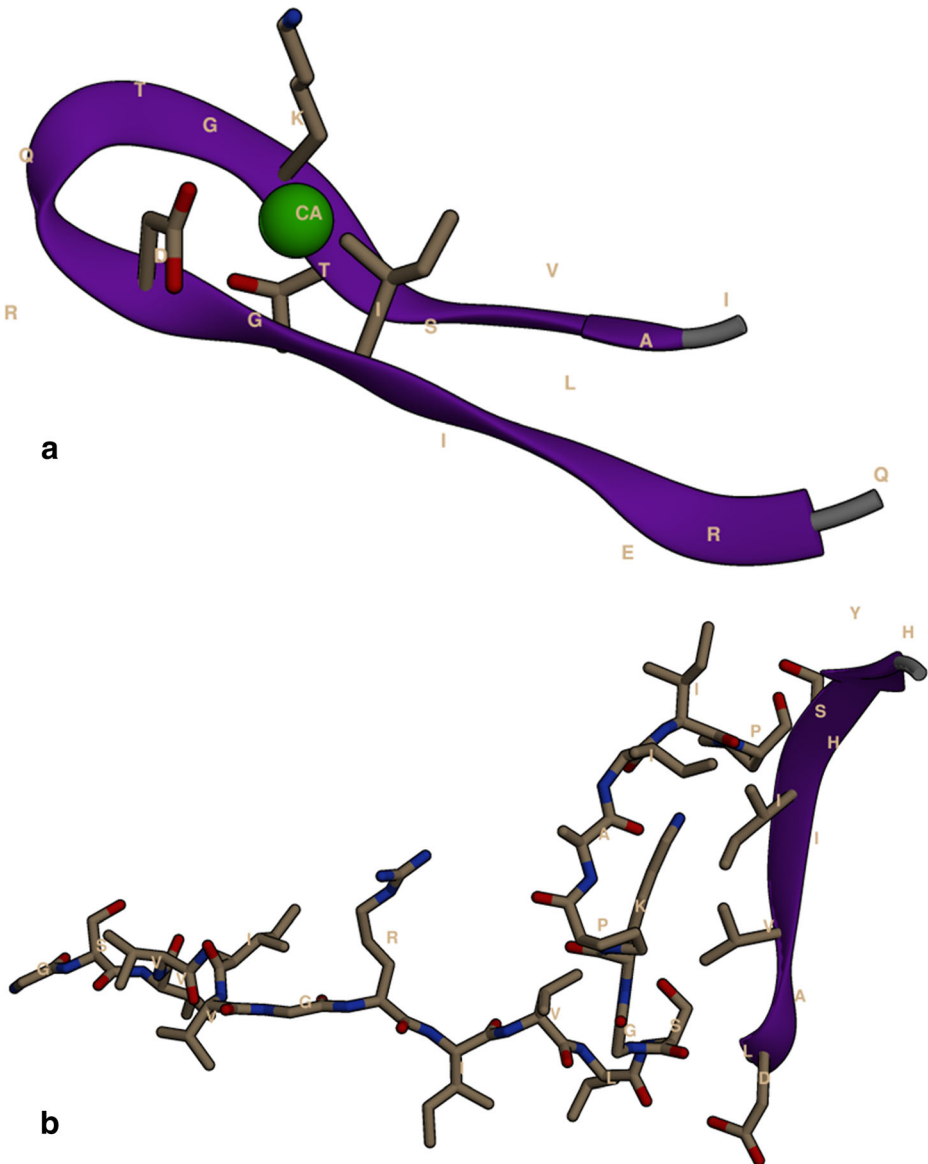
We emphasize that the fragments identified contain not only the amino acids of the ancestral phenotype, but they are the ones that prevail in diverse combinations, particularly if the amino acids G, A, D, V, S, T, N, or I participate in the binding of their prebiotic ligand. We illustrate our results with the structural reconstruction of the fragment recovered from the *ATP- $\alpha$* , and with the fragment encoded by RNY triplets in the *M $\beta$ L* family protein. Observe that the fragment of *ATP- $\alpha$*  can bind  $\text{Ca}^{2+}$  ( $C = 0.27$ , Fig. 1a), whereas the fragment of the *M $\beta$ L* can bind a random peptide ( $C = 0.29$ , Fig. 1b).

The data and structures of the peptides encoded by RNY triplets in *ATP- $\alpha$* , *rL6*, *DNA $\textit{topo}$* , *M $\beta$ L*, and *Hyp1*, together with the data of their corresponding randomisations, can be visualised in Table 1 (see also Supplementary Information II). Notice that the C-score values are in function of the structure and its probable ligand (Yang et al. 2013).

### Verification with Sequences and Structures

According to the consensus temporal order determined by Trifonov (2000), we noticed that most of the aa encoded by RNY are ancestral, therefore G, A, D, V, S, T, N, I are components of the earliest sequences (Trifonov et al. 2006). Sobolevsky and Trifonov (2006) derived octameric sequences, as well as their plausible structure based on its RMSD.

As can be seen in the trace depiction of the first octamer (Fig. 2a), of Sobolevsky and Trifonov (2006), we reconstructed their findings. Similarly, by following the same procedure with our own sequences, we also obtained the most probable ligand of that peptide (Fig. 2b). Data for three octamers, can be found in Supplementary Information II.



**Fig. 1** Structural reconstruction of two protein fragments encoded by RNY triplets, in bold their corresponding aa. In **a** the fragment recovered from the *ATP* -  $\alpha$ , which can bind  $\text{Ca}^{2+}$  ( $C = 0.27$ ) with the aa **G7**, **D8**, **T11**, **K13**, **T14**. In **b** the fragment encoded by RNY triplets of the *M $\beta$ L* family protein, which can bind a generic peptide ( $C = 0.29$ ) with the aa **D2**, **A3**, **V4**, **I5**, **I6**, **S7**

## Discussion

In this work, we reconstructed the first phenotype that would have been encoded by a primeval RNY code. Remarkably, the primeval phenotype did not encode for catalytic peptides, probably because catalysis was carried out by ribozymes or by small molecules that acted as electron shuttles (Gesteland et al. 2006), or even by clays such as montmorillonite (Jheeta and

**Table 1** Biological sequences

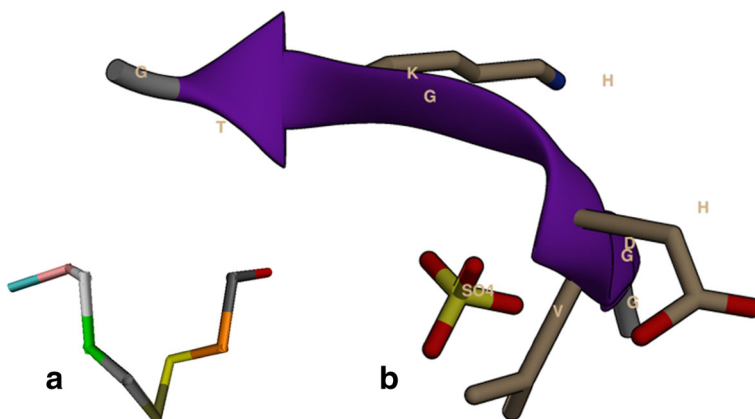
NAME	Ancestral aa/total length (original length)	C-score	Ligand	Involved aa
<i>ATP-<math>\alpha</math></i>	11/18 (18)	C = 0.27	Ca	<b>G7,D8,T11,K13,T14</b>
<i>rL6</i>	7/10 (9)	C = 0.20	SAH	<b>V5,T6</b>
<i>DNA<sub>topo</sub></i>	5/10 (10)	C = 0.60	Ca	<b>C4,P5,C7,Q8,K9</b>
<i>M<math>\beta</math>L</i>	6/10 (8)	C = 0.29	Peptide	<b>D2,A3,V4,I5,I6,S7</b>
<i>Hyp1</i>	8/15 (15)	C = 0.15	Ca	<b>E3,I5</b>

Summarised information concerning the examples: ATP-synthase  $\alpha$ -subunit(*ATP- $\alpha$* ), ribosomal protein L6 (*rL6*), DNA topoisomerase (*DNA<sub>topo</sub>*), a protein of the family metallo- $\beta$ -lactamases(*M $\beta$ L*), and one hypothetical protein (*Hyp1*). In the first column appears the name of the protein that contains a fragment codified by RNY triplets; the number of aa codified by RNY triplets and the total length of the fragment is indicated in the second column; the C-score for the structural models is on column three; columns 4 and 5 contain the probable ligand and the amino acids that bind it, respectively. aa coded by RNY triplets [G/A/D/V/S/T/N/I] is in bold

Joshi 2014). The peptides encoded by the primeval RNY genetic code acted as **cofactor stabilising binding sites, CSBSs**. It means that the first encoded peptides bound those molecules that have been proposed as prebiotic and essential for the development of life. Therefore, the collection of CSBSs can be called the primitive **bindome**, that turns out to be the **Ur-proteome**.

There are 16 triplets with RNY pattern and they code for 8 aa which can be synthesised prebiotically (all but Asn) according to Miller's experiments about primitive Earth conditions (Miller 1953; Lazcano and Miller 1996; Parker et al. 2011). On the other side, it has been demonstrated aptamers of RNA can bind aa to favour its polymerisation (Huang and Yarus 1997; Huang et al. 2000; Huang 2003).

The fragments retrieved contain amino acids coded by triplets other than RNY, albeit their respective ligands are mostly bound by G, A, D, V, S, T, N, or I. These fragments were reconstructed in 3D and their respective probable ligands were predicted. It is noteworthy that they bound preferably molecules considered ancestral and even prebiotic, such as those with a nucleotide moiety like NAD or S-adenosyl homocysteine SAH (Jadhav and Yarus 2002; Sharov 2016), small peptides randomly generated (Tamura and Schimmel 2003; Wieczorek



**Fig. 2** Reconstruction of the first octamer (GHVDHGKT) derived by Trifonov and Sobolevsky (2006). In **a** the structure in trace depiction is shown. In **b** the same octamer bound to  $\text{SO}_4$  (C = 0.37) with the aa H3, D5, G7, T9 is illustrated

et al. 2017) or generic nucleotide chains (Agmon 2016), or metallic ions such as  $Mg^{2+}$  or  $Ca^{2+}$  (Fedor 2002), or molecules like water or small organic molecules (Barton et al. 2007; Powner and Sutherland 2011).

On the other hand, validating our methodology and results, the structures reconstructed from the works of Trifonov, can also bind prebiotic ligands, precisely because some of its amino acids can be coded by RNY triplets. It is important to remember that possibly some peptides are not ancestral and are identified rather because they contain aa encoded by triplets RNY due to the accumulation of mutations, which do not detract the importance of this method in the identification of some primordia of the modern proteins.

To avoid over interpretations, we use the concept “cofactor” in its simplest meaning: a molecule that complexes with a protein and whose electrons will be transferred to the substrate during the catalytic process, it can be a metal ion or a coenzyme; thus the cofactor is transformed during reaction and it is essential to it (Nelson and Cox 2005). We are aware also about the well-known conception that in beginnings of life, the cofactors of ribozymes were nucleotide-cofactors such as NTPs as well as aa and the derivatives from both (Szathmáry 1993; Cech et al. 2006; Gesteland et al. 2006; Yarus 2011). However, it is more likely that those sequences were not the result of random mutations, because they are very highly conserved (Pagel and Pomiankowski 2007). Trifonov et al. (2006) and Sobolevsky and Trifonov (2006), suggested that those octamers could be older than LUCA and clearly most of its amino acids (G/A/D/V/S/T/N/I) are encoded by RNY triplets.

There are proposals different than the RNY as the first genetic code. For instance, Wong and Cedergren (1986) refused the possibility of existence of RNY vestiges. They argued that the bias towards RNY enrichment is due to natural selection with no reflection of ancestral coding. Ikehara (2009) proposed that genetic code began with pseudo-replication of the amino acids GADV, that corresponds to a GNC code, that turns out to be a subgroup of the RNY code, and it is the plausible source of the first polymerised peptides, just before the appearance of a recognizable genetic code. The code YRY(N)6YRY is also proposed as ancestral, given that it is a highly recurrent motif in the protein sequences of many organisms (Arqués et al. 1995). Other works have pointed out a coevolution between the metabolism and the genetic code (Wong 1975; Taylor and Coates 1989; Davis 1999; Di Giulio 2008).

However, the model based on the RNY as the first genetic code has a sound thermodynamic basis (Eigen and Schuster 1978; Eigen et al. 1985), reinforced by Shepherd’s observations (1981a, 1981b). Later, it was demonstrated that assuming the primeval RNY code, the SGC can be obtained by means of a sequence of symmetry breakings (José et al. 2007, 2009, 2011, 2015).

Several works (Koonin 2003; Mirkin et al. 2003; Delaye et al. 2005; Ranea et al. 2006; Sobolevsky and Trifonov 2006; Becerra et al. 2007; Itaya 1995; Mushegian and Koonin 1996; Koonin and Mushegian 1996; Fraser et al. 1995; Mushegian 1999; Hutchison et al. 1999; Gil et al. 2004; Glass et al. 2006; Sobolevsky et al. 2013), have proposed the protein content of LUCA, generally applying “top-down” approaches. On the other hand, with “bottom-up” methodologies (Miller 1953; White 1975; King 1980; Gilbert 1986; Wächtershäuser 1988, 2006; Schwartz 1995; Lazcano and Miller 1996; Miller et al. 1997; Huang and Yarus 1997; Huang et al. 2000; Segré et al. 2002; Harris and Christian 2003; Huang 2003; Chen et al. 2006; Janas et al. 2006; Joyce and Orgel 2006; Moore and Steitz 2006; Barton et al. 2007; Szathmáry 2007; Lazcano 2008; Sharov 2009; Lazcano 2010; Kua and Bada 2011; Kawamura 2012), investigators have delineated the first prebiological molecules. Notwithstanding that many of those scenarios are plausible it has not been possible to assemble a model that eventually would contain all the coincidental proto-biotic schemes.

More recently “top-down” and “bottom-up” approaches have been combined to reconstruct ancient proteins called *urzymes* (Li et al. 2011; Pham et al. 2010). The reconstructed “urzymes” are excellent examples of the most ancient catalytic sites in proteins. However, under the approach based on RNY triplets, the modules preceding the protein catalysis can be revealed because the scenario that we are dealing with implies that ribozymes performed the catalytic activities, whereas peptides solely bound the extant (prebiotic) molecules. Sobolevsky et al. (2013) revealed that the earliest protein modules were not involved in any catalytic activity. Indeed, the first tentative domains correspond with our results, given that those “omnipresent motifs” are indeed modules for cofactor binding.

Furthermore, the first encoded peptides made by G,A,D,V,S,T,N,I are in agreement with the SRM (Guimarães et al. 2008), in the sense that early encoded amino acids (i.e. the homogeneous sector) are typical of non-intrinsically organized protein secondary structures, that is, non-periodic, coils plus turns, non-alpha-helical plus non-beta-sheets, which are also dubbed ‘Intrinsically Disordered Protein Segments’. Therefore, the ‘informational patterns’ are not pre-existent but they are only potentials that will develop at the binding of the other interactants, in other words, the binding events. Based on our structural reconstructions, our CSBS are precisely barely organised, mostly they conform in coils with some portions resembling beta-sheets-like.

There is a database incorporating the different methodologies of massive genomes comparison using existing databases (UniProt, ByoCyc and KEGG) to reconstruct the nature of LUCA (Goldman et al. 2013). In such work, it is inferred that: 1) The RNA World hypothesis predicts that ribozymes carried out the most essential enzymatic functions, before the onset of translation 2) enzymatic functions requiring nucleotide cofactors, which may reflect the transition of a system based only on RNA to a system based on ribozymes and proteins 3) aa initially used as cofactors of ribozymes, may have played an early role in the transition to a system of protein enzymes; 4) cofactors FeS were also important from the start of biotic processes, because the mineral iron sulfide surfaces facilitated the production of small molecules and their polymerisation 5)  $Zn^{2+}$  cofactors could have been essential to the prebiotic processes to start nucleic acids chemistry and energy production. Our proposal is naturally reinforced in sense that cofactors began the transition from a prebiotic environment towards biotic processes. Then, it seems that the ancestral peptides constitute only one out of the five following modules: 1. Binding site for generic nucleotide chains 2. Binding site for molecules with a nucleotide moiety (e.g. ATP, NAD<sup>+</sup>, NADP, FMN, FAD, pyridoxal, CoA, S-AdoMet, SAH) 3. Binding site for ions and other small molecules 4. Binding site for generic peptides 5. Repetitive motifs with RNY-coded amino acids.

In effect our finding, about the CSBSs that emerged long before catalytic proteins, agrees with several notions about the very early origins of life because: 1. Moieties of some cofactors, such as the adenosine ring of ATP, NAD or FMN were abiotically synthesized and perhaps they were the only part that prevailed in those conditions, along with some metallic ions (Jadhav and Yarus 2002; White 1975; King 1980; Huang et al. 2000; Huang 2003; Szathmáry 1993; Sharov 2009; Yarus 2011). 2. Amino acids GADVSTNI, coded by RNY triplets, are characteristics from cofactor-binding motifs, but not from catalytic sites that rather contain amino acids such as H or K (as can be reviewed in the literature about each protein identified). 3. Usually, dependent-cofactor enzymes do not function without these molecules even in presence of the correct substrate, in some cases the enzyme can function if the substrate is present in huge quantities; hence, reasoning that evolution is not just about change but also about preserving the essential, the substrate is necessary but the cofactor is fundamental. 4.

Binding of molecules that were essentially electron carriers facilitated interactions among different prebiotic components; gradually some elementary catalytic reactions initiated and eventually they formed the metabolic cycles. 5. Lastly, the development of catalytic activities is preserved by natural selection; however, the processes at the dawn of life were essentially non-Darwinian but with gradual increase in complexity and cooperative instead of struggling (Eigen Eigen et al. (1989); Vetsigian et al. (2006); Goldenfeld and Woese (2007).

Hence, our method differs from the “bottom-up” and “top-down” approaches, because our rationale is based in the elucidation of the ancient genotype (RNY), extracted from modern organisms. In another mixed approach (bottom-up + top-down), Farias et al. (2016) reconstructed the ancestral tRNAs, they made concatamers only with those that currently charge the aa G,A,D,V,S,T,N,I; next they recovered, using BLASTX, the proteins with a portion encoded by those concatamers of proto-tRNAs. They found a distinct set of proteins than the one presented in this work, despite that both approaches assume an RNY code. We remark that both sets of peptides are not mutually exclusive. Most likely, both methods capture slightly distinct stages of evolution: our approach detects a somewhat earlier stage to that of Farias et al. (2016), because we are dealing with shorter RNY sequences. When we follow our present procedure but assuming Extended RNY codes (José et al. 2009), we found several matches like those found using tRNA concatamers (ongoing work). Alternatively, they could pertain to different populations of progenotes, as Woese envisioned (1998). Therefore, with the work of Farias et al. (2016) and with the present results we recovered different proto-modules that combined at the early origin of life to conform more complex peptides and eventually proteins.

Interestingly, the time necessary for the origin and evolution of life was not very long, and it fits nicely with Eigen’s statement: “the genetic code is not older than, but almost as old as our planet” (Eigen et al. 1989). Precisely the origin of life implies the origin of a genetic code, and not only has to do with random polymerisation of prebiotic molecules. The formation of the bindome(s) contributed to the evolutionary processes that occurred between prebiotic models and the evolution of the first forms of life. If life evolved in a Lamarckian way, assembling diverse modules and thus increasing complexity via HGT (Vetsigian et al. 2006; Goldenfeld and Woese 2007), beginning from some basic structures that could bind prebiotic and proto-biotic molecules, then the evolutionary processes would not have taken long time, making possible that carbon-based life emerged for the first time in this planet, including the raw material found in meteorites.

**Acknowledgements** Miryam Palacios-Pérez is a doctoral student from Programa de Doctorado en Ciencias Biomédicas (PDCB), Universidad Nacional Autónoma de México (UNAM) and she receives the fellowship 694877 from CONACYT. Fernando Andrade-Díaz is a doctoral student from PDCB, UNAM, and he receives the fellowship 204546 from CONACYT. Marco V. José was financially supported by PAPIIT-IN224015, UNAM, México. We thank Juan R. Bobadilla for his computer assistance. Molecular graphics and analyses for structures derived from the RYN model were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). Molecular visualisation of the structures reconstructed from Trifonov et al. (2006) was performed with VMD programme (<http://www.ks.uiuc.edu/Research/vmd/>).

## References

- Agmon I (2016) Could a proto-ribosome emerge spontaneously in the prebiotic world? *Molecules* 21:e1701  
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410

- Arqués D, Lapayre JC, Michel C (1995) Identification and simulation of shifted periodicities common to protein coding genes of eukaryotes, prokaryotes and viruses. *J Theor Biol* 172:279–291
- Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH (2007) Evolution. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
- Becerra A, Delaye L, Islas S, Lazcano A (2007) Very early stages of biological evolution related to the nature of the last common ancestor of the three major cell domains. *Annu Rev Ecol Evol Syst* 38:361–379
- Cech T (2012) The RNA Worlds in Context. New York, Cold Spring Harbor Laboratory Press. <https://doi.org/10.1101/cshperspect.a006742>.
- Cech TR, Moras D, Nagai K, Williamson JR (2006) The RNP world in The RNA world. In: Gesteland RF, Cech TR & Atkins JF (Eds). New York, Cold Spring Harbor Laboratory Press.
- Chen IA, Hanczyc MM, Szazani PL, Szostak JW (2006) protocells: genetic polymers inside membrane vesicles. In: Gesteland RF, Cech TR, Atkins JF (eds) The RNA World. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
- Davis BK (1999) Evolution of the genetic code. *Prog Biophys Mol Biol* 72:157–243
- Delaye L, Becerra A, Lazcano A (2005) The last common ancestor: what's in a name? *Orig Life Evol Biosph* 35: 537–554
- Di Giulio M (2008) An extension of the coevolution theory of the origin of the genetic code. *Biol Direct* 3:37
- Di Mauro E, Saladino R, Trifonov EN (2014) The path to life's origins. Remaining hurdles. *J Biomol Struct Dyn* 32(4):512–522
- Doolittle WF, Brown JR (1994) Tempo, mode, the progenote, and the universal root. *Proc Natl Acad Sci U S A* 91:6721–6728
- Eigen M, Schuster P (1978) The hypercycle, a principle of natural self-organization, part C: the realistic hypercycle. *Naturwissenschaften* 65:341–369
- Eigen M, Lindemann BF, Winkler-Oswatitsch R, Clarke CH (1985) Pattern analysis of 5S rRNA. *Proc Natl Acad Sci U.S.A.* 82:2437–2441
- Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress A, Haeseler A (1989) How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* 244:673–679
- Farias ST, Régio TG, José MV (2016) tRNA core hypothesis for the transition from the RNA world to the ribonucleoprotein world. *Life (Basel, Switzerland)* 6(2):e15
- Fedor MJ (2002) The role of metal ions in RNA catalysis. *Curr Opin Struct Biol* 12:289–295
- Forterre P, Philippe H (1999) The last universal common ancestor (LUCA), simple or complex? *Biol Bull* 196: 373–375
- Forterre P, Gribaldo S, Brochier C (2005) Luca: the last universal common ancestor. *Med Sci (Paris)* 21:860–865
- Fraser C, Gocayne J, White O, Adams M, Clayton R, Fleischmann R, Bult C, Kerlavage A, Sutton G, Kelley J, Fritchman R, Weidman J, Small K, Sandusky M, Fuhrmann J, Nguyen D, Utterback T, Saudek D, Phillips C, Merrick J, Tomb J, Dougherty B, Bott K, Hu P, Lucier T, Peterson S, Smith H, Hutchison C 3rd, Venter J (1995) The minimal gene complement of mycoplasma genitalium. *Science* 270:397–403
- Gesteland RF, Cech TR, Atkins JF (2006) The RNA world. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
- Gil R, Silva F, Peretó J, Moya A (2004) Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* 68:518–537
- Gilbert W (1986) The RNA world. *Nature* 319:618
- Di Giulio M. (2003) The universal ancestor and the ancestor of bacteria were hyperthermophiles. *J Mol Evol* 57: 721–730.
- Glansdorff N, Xu Y, Labedan B (2008) The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct* 9. <https://doi.org/10.1186/1745-6150-3-29>
- Glansdorff N, Xu Y, Labedan B (2009) The origin of life and the last universal common ancestor: do we need a change of perspective? *Res Microbiol* 160:522–528. <https://doi.org/10.1016/j.resmic.2009.05.003>
- Glass JL, Assad-Garcia N, Alperovich N, Yooshep S, Lewis M, Maruf M, Hutchison C 3rd, Smith H, Venter C (2006) Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A* 103:425–430
- Goldenfeld N, Woese C (2007) Biology next revolution. *Nature* 445:369
- Goldman AD, Bernhard TM, Dolzhenko E, Landweber LF (2013) LUCApedia: a database for the study of ancient life. *Nucleic Acids Res* 41:D1079–D1082. <https://doi.org/10.1093/nar/gks1217>.
- Guimarães RC, Moreira CH, Farias ST (2008) A self-referential model for the formation of the genetic code. *Theory Biosci* 127:249–270
- Harris M, Christian E (2003) Recent insights into the structure and function of the ribonucleoprotein enzyme ribonuclease P. *Curr Opin Struct Biol* 13:325–333
- Hoeningberg H (2003) Evolution without speciation but with selection: LUCA, the last universal common ancestor in Gilbert's RNA world. *Genet Mol Res* 4:366–375



- Huang F (2003) Efficient incorporation of CoA, NAD and FAD into RNA by —in vitro|| transcription. *Nucleic Acids Res* 31:e8
- Huang F, Yarus M (1997) Versatile 5' phosphoryl coupling of small and large molecules to an RNA. *Proc Natl Acad Sci U S A* 94:8965–8969
- Huang F, Bugg C, Yarus M (2000) RNA-catalyzed CoA, NAD, and FAD synthesis from phosphopantetheine, NMN, and FMN. *Biochemistry* 39:15548–15555
- Humphrey W, Dalke A, Schulten K (1996) VMD - visual molecular dynamics. *J Mol Graph* 14:33–38
- Hutchison CA, Peterson N, Gill R, Cline T, White O, Fraser C, Smith H, Venter C (1999) Global transposon mutagenesis and a minimal mycoplasma genome. *Science* 286:2165–2169
- Ikehara K (2009) Pseudo-replication of [GADV]-proteins and origin of life. *Int J Mol Sci* 10:1525–1537
- Itaya M (1995) An estimation of minimal genome size required for life. *FEBS Lett* 362:257–260
- Jadhav VR, Yarus M (2002) Coenzymes as coribozymes. *Biochimie* 84:877–888
- Janas T, Janas T, Yarus M (2006) RNA, lipids and membranes. In: Gesteland RF, Chech TR, Atkins JF (eds) *The RNA World*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
- Jheeta S, Joshi PC (2014) Prebiotic RNA synthesis by montmorillonite catalysis. *Life (Basel, Switzerland)* 4(3): 318–330
- José MV, Morgado ER, Govezensky T (2007) An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull Math Biol* 69:215–243
- José MV, Govezensky T, García JA, Bobadilla JR (2009) On the evolution of the standard genetic code: vestiges of critical scale invariance from the RNA world in current prokaryote genomes. *PLoS One* 4. <https://doi.org/10.1371/journal.pone.0004340>
- José MV, Morgado ER, Govezensky T (2011) Genetic Hotels for the Standard Genetic Code: evolutionary analysis based upon novel three-dimensional algebraic models. *Bull Math Biol* 73:1443–1476
- José MV, Zamudio GS, Palacios-Pérez M, Bobadilla JR, de Fariás ST (2015) Symmetrical and thermodynamic properties of phenotypic graphs of amino acids encoded by the primeval RNY code. *Orig Life Evol Biosph*. <https://doi.org/10.1007/s11084-015-9427-4>
- Joyce GF, Orgel LE (2006) Progress toward understanding the origin of the RNA world. In: Gesteland RF, Chech TR, Atkins JF (eds) *The RNA World*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
- Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) KEGG from genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34: D354–D357
- Kawamura K (2012) Drawbacks of the ancient RNA-based life-like system under primitive earth conditions. *Biochimie* 94:1441–1450
- Kim KM, Caetano-Anollés G (2011) The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evol Biol* 11. <https://doi.org/10.1186/1471-2148-11-140>
- King GAM (1980) Evolution of the coenzymes. *Biosystems* 13:23–45
- Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 1:127–136
- Koonin E, Mushegian AR (1996) Complete genome sequences of cellular life: glimpses of theoretical evolutionary genomics. *Curr Opin Genet Dev* 6:757–762
- Kua J, Bada JL (2011) Primordial ocean chemistry and its compatibility with the RNA world. *Orig Life Evol Biosph* 41:553–558
- Lazcano A (2008) What is life? A brief historical overview. *Chem Biodivers* 5:1–15
- Lazcano A (2010) Historical development of origins research. *Cold Spring Harb Perspect Biol* 2:a002089
- Lazcano A, Miller SL (1996) The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. *Cell* 85:793–798
- Lehman J (2002) Amplification of the sequences displaying the pattern RNY in the RNA world: the translation → translation/replication hypothesis. *J Theor Biol* 219:521–537
- Li L, Weinreb V, Francklyn C, Carter C Jr (2011) Histidyl-tRNA synthetase urzymes: class I and II aminoacyl tRNA synthetase urzymes have comparable catalytic activities for cognate amino acid activation. *J Biol Chem* 286:10387–10395
- Miller SL (1953) A production of amino acids under possible primitive earth conditions. *Science* 117:528–529
- Miller S, Schopf J, Lazcano A (1997) Oparin's "origin of life": sixty years later. *J Mol Evol* 44:351–353
- Mirkin B, Fenner T, Galperin M, Koonin E (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3:2
- Moore PB, Steitz TA (2006) The roles of RNA in the synthesis of protein. In: Gesteland RF, Chech TR, Atkins JF (eds) *The RNA World*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
- Morange M (2011) Some considerations on the nature of LUCA, and the nature of life. *Res Microbiol* 162:5–9. <https://doi.org/10.1016/j.resmic.2010.10.001>

- Mushegian A (1999) The minimal genome concept. *Curr Opin Genet Dev* 9:709–714
- Mushegian A (2008) Gene content of LUCA, the last universal common ancestor. *Front Biosci* 13:4657–4666
- Mushegian A, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *PNAS* 93:10268–10273
- Nelson D, Cox M (2005) *Lehninger principles of biochemistry*, 4th edn. W.H. Freeman and Company, USA
- Pagel MD, Pomiankowski A (2007) Evolutionary genomics and proteomics. Sinauer Associates, USA
- Parker ET, Cleaves HJ, Dworkin JP, Glavin DP, Callahan M, Aubrey A, Lazcano A, Bada JL (2011) Primordial synthesis of amines and amino acids in a 1958 miller H<sub>2</sub>S-rich spark discharge experiment. *Proc Natl Acad Sci U S A* 108:5526–5531
- Penny D, Poole AM (1999) The nature of the last universal common ancestor. *Curr Opin Genet Dev* 9:672–677
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605–1612
- Pham Y, Kuhlman B, Butterfoss G, Hu H, Weinreb V, Carter C Jr (2010) Tryptophanyl-tRNA synthetase Urzyme: a model to recapitulate molecular evolution and investigate intramolecular complementation. *J Biol Chem* 285:38590–38601
- Poole A, Logan DT (2005) Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA genome? *Mol Biol Evol* 22:1444–1455
- Powner MW, Sutherland JD (2011) Prebiotic chemistry: a new modus operandi. *Philos Trans R Soc Lond B* 366(1580):2870–2877
- Ranea J, Sillero A, Thornton M, Orengo A (2006) Protein superfamily evolution and the last universal common ancestor (LUCA). *J Mol Evol* 63:513–525
- Schwartz AW (1995) The RNA world and its origins. *Planet Space Sci* 43:161–165
- Segré D, Ben-Eli D, Deamer D, Lancet D (2002) The lipid world. *Orig Life Evol Biosph* 31:119–145
- Sharov A (2009) Coenzyme autocatalytic network on the surface of oil microspheres as a model for the origin of life. *Int J Mol Sci* 10:1838–1852
- Sharov AA (2016) Coenzyme world model of the origin of life. *Biosystems* 144:8–17
- Shepherd JC (1981a) Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. *J Mol Evol* 17:94–102
- Shepherd JC (1981b) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc Natl Acad Sci U. S. A.* 78:1596–1600
- Sobolevsky Y, Trifonov E (2006) Protein modules conserved since LUCA. *J Mol Evol* 63:622–634
- Sobolevsky Y, Guimarães R, Trifonov E (2013) Towards functional repertoire of the earliest proteins. *J Biomol Struct Dyn* 31:1293–1300
- Szathmáry E (1993) Coding coenzyme handles: a hypothesis for the origin of the genetic code. *Proc Natl Acad Sci U S A* 90:9916–9920
- Szathmáry E (2007) Coevolution of metabolic networks and membranes: the scenario of progressive sequestration. *Philos Trans R Soc B* 362:1781–1787
- Tamura K, Schimmel P (2003) Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptors. *PNAS* 100:8666–8669
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4. <https://doi.org/10.1186/1471-2105-4-41>.
- Taylor FJ, Coates D (1989) The code within the codons. *Biosystems* 22:177–187
- Trifonov E (2000) Consensus temporal order of amino acids and evolution of the triplet code. *Gene* 261:139–151
- Trifonov E (2006) Self-inflicted fear of evolution. *Orig Life Evol Biosph* 36:557–558
- Trifonov EN, Sobolevsky Y (2006) Protein modules conserved since LUCA. *J Mol Evol* 63:622–634
- Trifonov E, Gabdank I, Barash D, Sobolevsky Y (2006) Primordia vita, deconvolution from modern sequences. *Orig Life Evol Biosph* 36:559–565
- Vetsigian K, Woese C, Goldenfeld N (2006) Collective evolution and the genetic code. *Proc Natl Acad Sci U S A* 103:10696–10701
- Wächtershäuser G (1988) Before enzymes and templates: theory of surface metabolism. *Microbiol Rev* 52:452–484
- Wächtershäuser G (2006) From volcanic origins of chemoautotrophic life to bacteria, Archaea and Eukarya. *Phil Trans R Soc B* 361:1787–1808
- White HB 3rd (1975) Coenzymes as fossils of an earlier metabolic state. *J Mol Evol* 7:101–104
- Wieczorek R, Adamala K, Gasperi T, Polticelli F, Stano P (2017) Small and random peptides: an unexplored reservoir of potentially functional primitive organocatalysts. The case of seryl-histidine. *Life (Basel, Switzerland)* 7(2):E19

- Woese CR (1998) The universal ancestor. *Proc Natl Acad Sci U S A* 95:6854–6859
- Woese CR (2000) Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A* 97:8392–8396
- Wong J (1975) A co-evolution theory of the genetic code. *Proc Natl Acad Sci U S A* 72:1909–1912
- Wong J, Cedergren R (1986) Natural selection versus primitive gene structure as determinant of codon usage. *Eur J Biochem* 159:175–180
- Wong J, Chen J, Mat W, Ng K, Xue H (2007) Polyphasic evidence delineating the root of life and roots of biological domains. *Gene* 403:39–52
- Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J* 101:2525–2534
- Yang CM (2005) On the structural regularity in nucleobases and amino acids and relationship to the origin and evolution of the genetic code. *Orig Life Evol Biosph* 35:275–295
- Yang J, Roy A, Zhang Y (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 29:2588–2595
- Yarus M (2002) Primordial genetics: phenotype of the ribocyte. *Annu Rev Gen* 36:125–151
- Yarus M (2011) Getting past the RNA world: the initial Darwinian ancestor. *Cold Spring Harb Perspect Biol* 3: a003590. <https://doi.org/10.1101/cshperspect.a003590>
- Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:40



# The evolution of proteome: From the primeval to the very dawn of LUCA

Miryam Palacios-Pérez, Marco V. José\*

Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Ciudad de México CDMX, C.P. 04510, Mexico



## ARTICLE INFO

### Keywords:

RNY code  
Ur-proteome  
Extended genetic codes I and II  
Extended proteomes  
Progenotes  
FUCA  
LUCA  
Evolution  
Of  
Genetic code

## ABSTRACT

The attempt to delineate the essential features that characterized life in its beginnings and the understanding of how those features evolved, represent important scientific challenges. While there have been varied efforts in the elucidation of how the first biomolecules arose from a prebiotic environment, there remains important gaps towards the characterization of the complete repertoire of the Last Universal Common Ancestor (LUCA). We portray a step-wise proteome evolution, by looking at the phenotype encoded by each one of the genetic codes that were appearing along evolution, beginning with the primeval genetic code and then with two intermediate Extended RNA codes, which finally shaped the current Standard Genetic Code (SGC). Notably, all molecules involved in translation, such as ribosomal proteins and all tRNA synthetases, were already present before LUCA. The metabolism belonged to extremophiles as hinted by the presence of reverse gyrase and acetyl coenzyme A synthase. Furthermore, we predict the structure and possible ligands of the proteins retrieved. We have forged a bridge between the hitherto unknown proteome of progenotes and the proteome of LUCA.

## 1. Introduction

Extensive analyses have been made to reconstruct the transition from pre-biotic to proto-biotic in one side (e.g. Miller, 1953; Wächtershäuser, 1988; Johnson et al., 2008; Sharov, 2009; Cech, 2012; Szathmáry, 2007) and to uncover the genetic composition of the Last Universal Common Ancestor (LUCA) of cells (e.g. Mushegian, 1999; Koonin, 2003; Glansdorff et al., 2008; Forterre et al., 2005; Delaye et al., 2005; Ouzounis et al., 2006; Weiss et al., 2016) in the other side. To advance in solving these fundamental questions, there have been some notable experiments of ancestral sequence reconstructions (Thornton, 2004; Merkl and Sterner, 2016) or paleoenzymology (Carter, 2014). However, the events that occurred between the first forms of life to the appearance of LUCA, remain poorly understood. The appearance of LUCA is the result of several evolutionary episodes, each one characterized by different fundamental components: Prebiotic molecules → RNA + amino acids / peptides → RNA / Proteins → RNA / Proteins + DNA → DNA / Proteins + RNA, where this last phase corresponds to the content of LUCA (cf. Cech, 2012).

Eigen and Schuster (1978) proposed that the initial genetic code, corresponding to the RNA world, started with ribonucleotide chains of concatenated triplets with RNY (purine-any base-pyrimidine) pattern (where R = [A/G], Y = [C/U] and N = [A/C/G/U], following in all cases the parity rules R:Y). The symmetric and thermodynamic properties of the phenotypic graphs of the amino acids (aa) encoded by such ancestral RNY code have been reported (José et al., 2015). Using algebraic models, it has been demonstrated that it is possible to generate the standard genetic code (SGC) by means of two different evolutionary pathways starting from the primeval RNY code (José et al., 2007, 2011). The Extended RNA code type I (Ex1), is obtained by changes of reading frame of RNY chains, and the Extended RNA code type II (Ex2), is achieved by transversions of the first or the third nucleotide. Ex1 consists of 16 codons of the type RNY plus 32 codons obtained by considering the RNA code but in the second (NYR type) and third (YRN type) reading frames. Ex2 comprises all codons of the type RNY plus 32 codons that arise from transversions of the RNA code in the first (YNY type) and third (RNR) nucleotide bases. Both routes converge and complement the missing triplets in one extended RNA code with the

**Abbreviation:** aaRL, aminoacyl-tRNA ligases; aaRS, amino-acyl-tRNA synthetases; AcCoA, acetyl-coenzyme A; AcCoAs, AcCoA synthase; C-score, Confidence score; CDS, Coding sequences; CSBS, Cofactor Stabilising Binding Sites; CW, coding-wise; EGCs, Extended Genetic Codes; EPh, Extended Phenotype; Ex1, Extended RNA code type I; Ex2, Extended RNA code type II; ExP, Extended Proteomes; LUCA, Last Universal Common Ancestor; MSA, Multiple Sequence Alignment; ORF, Open reading frame; PGC, Primeval Genetic Code; PPh, Primeval Phenotype; reGyr, reverse gyrase; SGC, Standard Genetic Code; TM-score, Template modelling score; ThrRL, Threonine-tRNA ligase; TPI, triose-phosphate isomerase; TrpRL, tryptophan-tRNA ligase; Amino acids, Glycine (G), Alanine (A), Aspartic acid (D), Valine (V), Proline (P), Serine (S), Glutamic acid (E), Leucine, Threonine (L, T), Arginine (R), Isoleucine, Glutamine, Asparagine (I, Q, N), Histidine (H), Lysine (K), Cysteine (C), Phenylalanine (F), Tyrosine (Y), Methionine (M), Tryptophan (W)

\* Corresponding author.

E-mail addresses: [mir.pape@iibiomedicas.unam.mx](mailto:mir.pape@iibiomedicas.unam.mx) (M. Palacios-Pérez), [marcojose@biomedicas.unam.mx](mailto:marcojose@biomedicas.unam.mx) (M.V. José).

<https://doi.org/10.1016/j.biosystems.2019.04.007>

Received 26 February 2019; Received in revised form 9 April 2019; Accepted 10 April 2019

Available online 14 April 2019

0303-2647/ © 2019 Elsevier B.V. All rights reserved.

triplets of the other extended RNA code. The genomes obeying either the Ex1 or Ex2, correspond to the progenotes proposed by Woese (Woese, 1998). It has been shown that the codons of current genomes, and those that contained only RNY (16 triplets), or Ex1 (48 triplets), or Ex2 (48 triplets), show critical scale invariance (José et al., 2009), i.e. the statistical properties that characterize the distance series of each codon were fixed in each code, including the SGC, and organisms have evolved under the same fundamental rules for at least 3.5 billion years. We emphasize that the symmetrical model of the evolution of the SGC (José et al., 2007, 2011, 2017) is completely equivalent to the Rodin-Ohno model that divides the table of the SGC into 2 classes of aminoacyl-tRNA synthetases (aaRS) (Rodin and Ohno, 1997). The SGC, as derived from the primeval genetic code through symmetry breakings, and the Rodin-Ohno model are one and the same (José et al., 2017). The rationale of our approach is that if a present-day long genome shares a vital characteristic of its theoretical shorter earlier self then one knows something about its ancestor, and by extension the common ancestor of its relatives. Woese (1998) contended that by exploiting this approach further and examining genomes closer, one may gain a deeper understanding of our universal ancestor and how the SGC could have originated. Classical examples of this approach are the seminal works of Eigen and Winkler-Oswatitsch (1981a,b) about the reconstruction of early tRNA and a primordial gene in which they showed that nucleotide sequences existing today still carry, in a hidden form, the patterns of their ancestry.

To shed light on the intermediate evolutionary stages, we looked for the phenotypes encoded in ancient genetic codes, beginning with the primeval genetic code which encoded Cofactor Stabilising Binding Sites (CSBS) (Eigen and Schuster, 1978; Shepherd, 1981a,b, Lehmann, 2002, José et al., 2015; Palacios-Pérez et al., 2018), and subsequently we generated the genomes corresponding to the intermediate Extended Genetic Codes (Ex1 and Ex2) (José et al., 2009, 2011). To depict the evolution of the proteome, we gathered the common proteins at each stage which coincided with many proteins delineated as ancient by other independent recent works (Farias et al., 2016; Weiss et al., 2016; Lupas and Alva, 2017). Furthermore, we predict the structures and probable ligands of some retrieved common proteins. In summary, our findings fill, in part, the gap in understanding the evolution of proteomes since the first forms of life to the proteomes of progenotes just before the emergence of LUca.

## 2. Methods

From each common protein we found different fragments and some of them were concatenated orderly. The concatenated protein fragments from each organism were aligned to obtain a consensus, and their structure and possible ligands were predicted.

### 2.1. Data sources

We retrieved the genes in CDS, from <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/> (as \*cds\*.fna\*) of 13 bacteria and 13 archaea: *Aquifex aeolicus* VF5 (AqfxV, NC\_000918.1), *Bacillus subtilis* 168 (Basub, NC\_000964.3), *Borrelia burgdorferi* B31 (Bobur, NC\_001318.1), *Deinococcus radiodurans* R1 (Derad, NC\_001263.1 y NC\_001264.1), *Escherichia coli* K12 MG1655 (EcoK12, NZ\_CP025268.1), *Mycoplasma genitalium* G37 (Mygen, NC\_000908.2), ca. *Pelagibacter ubique* HTCC1062 (Peubi, NC\_007205.1), *Shewanella piezotolerans* WP3 (Shpz3, NC\_011566.1), *Streptococcus agalactiae* A909 (SagA, NC\_007432.1), *Synechococcus* CC9902 (SynCC, NC\_007513.1), *Thermotoga maritima* MSB8 (Thmar, NC\_000853.1), *Thermus aquaticus* Y51MC23 (TaqY51, NZ\_CP010822.1), *Thermus thermophilus* HB8 (Ther2, NC\_006461.1); *Acidianus hospitalis* W1 (Aciho, NC\_015518.1), ca. *Nitrosopumilus sediminis* AR2 (Nised, NC\_018656.1), *Haloarcula marismortui* ATCC 43049 (Hamar, NC\_006396.1 y NC\_006397.1), *Haloferax volcanii* DS2 (Hxvol, NC\_013967.1), *Haloquadratum walsbyi* DSM 16790 (Haqwa,

NC\_008212.1), *Korarchaeum cryptofilum* OPF8 (Kocry, NC\_010482.1), *Methanocaldococcus jannaschii* DSM\_2661 (Mejan, NC\_000909.1), *Methanosarcina acetivorans* C2A (Macet, NC\_003552.1), *Pyrococcus furiosus* DSM 3638 (Pyfur, NC\_003413.1), *Sulfolobus acidocaldarius* DSM 639 (Sacid, NC\_007181.1), *Thermococcus gammatolerans* EJ3 (Thgam, NC\_012804.1), *Thermococcus sibiricus* MM739 (Thsib, NC\_012883.1), *Thermoplasma volcanium* GSS1 (Tmvol, NC\_002689.2).

### 2.2. Reconstruction of ancient genomes

To reconstruct the original structure of the genome, all ORFs were assembled one after the other, i.e. coding-wise (CW) with an *ad hoc* programme, that deletes the headers and therefore concatenates the genes one after the other, just as they are reported in the file \*cds\*.fna\*, making easier its comparison with a posterior alignment approach; as negative control, a random sequence using the CW genome was generated, shuffling thrice the nucleotides to eliminate the biological sense of the sequence. From the CW sequence and from its corresponding shuffled control we used three Perl scripts to discard the triplets that do not belong to RNY or Extended Codes, Ex1 and Ex2.

The RNY code was shown to be already frozen (José et al., 2015), whilst the Extended RNA codes and the SGC emerged from symmetry breakings (José et al., 2009; José et al., 2011). As mentioned above, the scaling properties of the distance series of some codons from the RNA code and most codons from both extended RNA codes are identical or very close to the scaling properties of codons of the SGC (José et al., 2009). The latter implies that the sequences analysed in the present work, are vestiges or fossilized remnants of the primeval and the extended RNA codons in today existing genes. Therefore, the assignments of codons of amino acids since the RNY code are assumed to be the same as the current SGC.

According to Eigen and Schuster (1978), RNY is the only pattern thermodynamically consistent because it presents self-complementarity with no excess of purines or pyrimidines, and it is capable of evolve through hypercycles. Additionally, those triplets encode the most abundant amino acids prebiotically synthesised (Miller, 1953). There are other patterns that have self-complementarity like YNR, however, such type of triplets is less abundant than RNY in biological genomes (Eigen et al., 1985); besides, all the three stop-codons UGA, UAA, and UAG, are YNR and they have been recognised as late-comer triplets in most of the proposals (Trifonov, 2004). Nonetheless, we performed a positive control with the 26 organisms, preserving only YNR triplets that was shuffled three times to establish a proper cut-off value. We found that for most organisms, the retrieved proteins with YNR are fewer than those found with RNY and the corresponding cut-off E-value is one order of magnitude greater than the one obtained with the primeval genetic code. For instance, the E-value for RNY-encoded proteins of “Derad” begins in 2.00E-34 and the cut-off is established in 0.002 which encompasses 115 unique proteins; the E-value for YNR-encoded proteins starts in 0.02 with a cut-off set in 0.025 which clearly leads to any protein in the list.

### 2.3. Proteomes retrieved

We compared the biological genomes, the primeval and both extended, and the controls of each one with the original coding sequences of the file \*cds\*.fna\* of every organism, using BLASTn in its installed version (Altschul et al., 1990). Given that RNY and both Extended Codes possess less triplets than the SGC, and, bearing in mind our premise that the former evolutionary stages were earlier than the latter, parameters were adjusted to allow as most outcomes as possible, than regularly are used in BLAST searches, only preserving the maximum E-value at 10. Then, to determine a cut-off value to the retrieved proteins, we made numerical comparisons of the E-values of each of the biological primeval and extended genomes and their corresponding controls, the cut-offs were then set per each out of the three genomes, for each

organism.

Triplets of the type RNY, Ex1 or Ex2 encode several fragments of each protein, hence, to simplify the qualitative extraction of information we changed the headers with an *ad hoc* programme, that leaves the sequence intact but change the header to preserve only the name of the protein, deleting GenBank identifiers, the position in genome and some other additional information. Then we recorded one time each result per organism to make comparisons among the 26 organisms, and we computed as *common proteins* all that are present in at least two organisms. The common proteins were classified according to its function, based on COG (Tatusov et al., 2003), KEGG (Kanehisa et al., 2006), and EggNOG (Huerta-Cepas et al., 2016) databases, as well as pertinent literature for some entries.

#### 2.4. Assembling and alignments of ancient proteins

Seven out of all common proteins were selected to illustrate the relevance of this approach to explain evolutionary traits. First, we translated the fragments of each of those 7 proteins with the EMBOSS program "sixpack" (Li et al., 2015, [http://www.ebi.ac.uk/Tools/st/emboss\\_sixpack/](http://www.ebi.ac.uk/Tools/st/emboss_sixpack/)); subsequently, we assembled those translated fragments orderly and we made multiple sequence alignments per protein using M-coffee webserver, because it combines several alignment methods which ensures highly accurate outcomes (Moretti et al., 2007); the MSA output was visualised in UGENE software (Okonechnikov et al., 2012) to generate a consensus sequence from which the structure and possible ligands could be deduced. In most of the selected proteins the portions encoded by RNY triplets were completely different, therefore one consensus sequence per protein could not be generated and each moiety was treated individually. For instance, if fragments GSEIFKATHERSS and MVIAMIPAAF are encoded by Ex2 triplets, according with "BLASTn" and then translated with "sixpack", but the original protein contains the sequence MVIAMIPAAFSGSEIFKATHERSS, then those fragments are assembled as: MVIAMIPAAFxxxxxGSEIFKATHERS, and the final sequence will be conformed as: MVIAMIPAAFSGSEIFKATHERSS..., which will be used to generate the MSA.

#### 2.5. Structural prediction of ancient proteins

To verify the functionality of generated consensus, we introduced them sequentially into pertinent on-line programmes of the suite Zhang Lab (<https://zhanglab.cmb.med.umich.edu/>), that allow to non-experts to generate structural models using the own sequences through Internet services. The first approximation was made with I-TASSER programme (Zhang, 2008), that is oriented to reconstruct three dimensional protein structures given a FASTA sequence; I-TASSER is a hierarchical protein structure modeller that first identify known templates, experimentally determined, whose sequence is similar to portions of the FASTA sequence provided, 3D fragments are assembled and the portions that remain unpredicted because there are not templates are then modelled *ab initio*, structure is reassembled taking into account spatial and thermodynamic restrictions, the structure with lowest energy is finally selected. The best output model was relaxed with ModRefiner programme (Xu and Zhang, 2011) in order that the side-chain and backbone atoms be more flexible and not restrained by the reference model, avoiding steric clashes. The probable ligands of the refined structures were predicted with the meta-server COACH (Yang et al., 2013a), an approach for protein-ligand binding site prediction based on comparison of the structure provided by the user with the experimental ligand-binding templates, the predictions generated are compared with results of other methodologies to generate final ligand binding site predictions, ranked from the highest to the lowest C-score; the C-score, or confidence score, summarises the probability that a given structure binds the proposed ligand (abbreviated according with BioLip database: Yang et al., 2013b, <https://zhanglab.cmb.med.umich.edu/BioLIP/ligand.html>), i.e. the more reliable prediction given that it

is supported by the largest number of templates.

To identify which fragment is encoded by each genetic code associated to the modern protein, we selected some experimentally determined structures, preferentially, but not exclusively, from at least one out of the 26 organisms included in our set; TM-align (Zhang and Skolnick, 2005) was used to structurally align the peptides encoded by RNY triplets to the peptides encoded by each genetic code, with the respective crystallographic structures downloaded from <http://www.rcsb.org/>; the quality in alignment, i.e. how similar is our predicted structure encoded by some ancient genetic code with respect to the complete modern protein experimentally determined, is indicated by the TM-score, or template modelling score that always lies between the interval (0,1] with better templates having higher TM-score, which is a metric used for measuring the similarity of two protein structures, residue per residue, no matter the difference in length of both structures; according with statistics (Zhang and Skolnick, 2005),  $0.0 < \text{TM-score} < 0.30$  indicates random structural similarity, whereas  $0.5 < \text{TM-score} < 1.00$  indicates that both structures are in about the same fold, leaving the range  $0.31 < \text{TM-score} < 0.49$  open to the interpretation according with users data. Assemblies were finally visualised with Chimera (Pettersen et al., 2004).

### 3. Results

RNY triplets encode 7 out of 10 amino acids (aa) synthesised in Miller-type experiments (Miller, 1953; Johnson et al., 2008), while the aa encoded whether by E1 or by E2, consistently appear later in the evolution of the genetic code (Trifonov, 2004). Bearing in mind that "a conservation degree of a protein sequence is virtually proportional to its antiqueness" (Pagel and Pomiankowski, 2008), we searched for the phenotype encoded by each genetic code, using the genomes of 26 organisms –13 archaea and 13 bacteria– phylogenetically unrelated among them, whence we obtained the common proteins, retrieving thus the phenotypes corresponding to each genotype. In this way, RNY triplets encode the Ur-proteome and both **extended genetic codes** (EGCs) encode **extended proteomes** (ExP) that would constitute the **extended phenotype** (EPH) of life, depicting thus the evolution of proteomes since its very origins up to the very dawn of LUca.

#### 3.1. Primeval peptides constitute the Ur-proteome

When we searched for the phenotype encoded by the **primeval genetic code** (PGC), we obtained short lists of proteins, corresponding to the **primeval phenotype** (PPH), whose portions were encoded by RNY triplets. Even though we obtained only short fragments out of few proteins (less than 200) after comparing the PPH of 26 organisms analysed (13 archaea and 13 bacteria, phylogenetically distantly related), such Ur-proteome agrees with what we previously obtained with only one genome (Palacios-Pérez et al., 2018) and that corresponds with CSBs or *bindome*. From the 26 organisms, we found that common proteins with one fragment encoded by RNY triplets pertain currently to diverse cellular processes, to wit (Table 1): from the translation process we found some ribosomal proteins, aminoacyl-tRNA synthetases (aaRS) (now known as aminoacyl-tRNA ligases (aaRL) <https://www.uniprot.org/keywords/KW-0030>), some translation factors, proteins involved in RNA binding, RNA processing, and RNA modification, and some proteins involved in ribosome biogenesis; some other proteins are implicated in replication, or processing, or repairing, or modification of DNA; some metabolic proteins also contain regions encoded by RNY triplets, such as fructose 1,6-bisphosphatase or triose-phosphate isomerase; we also found some proteins for the cell-wall assembly, for the transport of diverse molecules, cell signalling proteins, and even some proteins involved in flagellum synthesis (Table 1).

Recalling that not the whole proteins identified are ancient, but portions of the genes that encode them are made of RNY triplets, it is remarkable that most of those proteins are among proteins stated as the

**Table 1**  
Non-exhaustive but representative list of proteins with RNY-encoded regions.

Few subunits of some topoisomerases and gyrases
Few subunits of DNA polymerase
RNA modification enzymes, SAM-dependant
Few tRNA ligases (aaRS)
Few ribosomal proteins
Few factors of translation apparatus (such as EF-G)
Some metabolic proteins of energetic metabolism, particularly electron transfers (ferredoxins, cytochrome c, or CoA-dependant enzymes)
Some few enzymes of glycolysis (such as fructose 1-6-bisphosphatase)
Some few enzymes of amino acids metabolism (such as 3-phosphoshikimate 1-carboxyvinyltransferase)
Some few enzymes of nucleotides metabolism (such as CTP synthetase)
Some few enzymes of lipids metabolism (such as long-chain-fatty-acid-CoA ligase)
Few peptidases and proteases
Chaperones and chaperonins
Few enzymes for synthesis and modification of membrane proteins (such as glycosyltransferases)
Some ABC transporters components
Other transporters (such as few components of PTS, antiporters, and channels (such as porin))
Kinases (Ser/Thr, Asp)
Some subunits of proteins for flagellar mobility
Few cell division proteins
Multiple functions / unknown functions
DUFs
Hypothetical proteins

most ancient (cf. [Farias et al., 2016](#); [Lupas and Alva, 2017](#)), in particular those related with the translation process and RNA modifications; even if some other proteins identified in this evolutionary stage do not make sense, as is the case of flagellar components, note that in all cases the ancient portion is a CSBS. The CSBSs constitute only one out of the five following modules: 1) binding site for generic nucleotide chains, 2) binding site for molecules with a nucleotide moiety (e.g. ATP, NAD<sup>+</sup>, NADP, FMN, FAD, pyridoxal, CoA, S-AdoMet, SAH), 3) binding site for ions and other small molecules, 4) binding site for generic peptides, 5) repetitive motifs with RNY-coded amino acids ([Palacios-Pérez et al., 2018](#)). That implies that the CSBSs acted as the first building blocks of proteins, which later would evolve and diversify into more modern conformations encoded by more complex genetic codes, as it will be illustrated in concrete examples of next sections.

### 3.2. Extended phenotypes

After comparing the proteins encoded by Ex1 and Ex2 of the 26 organisms, it seems that both EGCs encode nearly the same proteins, which pertain to the most diverse cellular processes. It is noteworthy that many of the ribosomal proteins and all the aaRS are part of both Ex1 and Ex2 sets, as well as the translation factors and a high number of RNA processing enzymes such as RNA helicases and the DNA primase (that is actually an RNA polymerase), and RNA modification proteins, whether for rRNAs, tRNAs or generic ribonucleic chains, including the enzyme that adds the CCA terminus to tRNAs. All subunits of the ATP synthases (F<sub>0</sub>F<sub>1</sub> and V-type) also contain portions encoded by both EGCs, as well as cytochromes and many other subunits of the oxidative phosphorylation chain; several other proteins of the diverse energetic metabolisms, such as carbon fixation, sulphur, methane, and nitrogen metabolism have also portions encoded, in part, by Ex1 and Ex2; many other housekeeping proteins are also encoded, in part, by Ex1 and Ex2 triplets, such as proteins involved in metabolism of carbohydrates, lipids, nucleotides, and amino acids. Other proteins, partially encoded by Ex1 and Ex2, are involved in cell wall or external membrane synthesis (peptidoglycan, lipoteichoic acid), or the synthesis of membrane decorations complexes (glycoproteins, lipoproteins, glycol-lipoproteins, and lipo-polysaccharides). Fragments of several ABC transporters are encoded by EGCs, as well as other type of transporters and channels,

such as antiporters, symporters, MFS transporters, or porins. Proteins that modify other proteins contain also regions encoded by Ex1 and Ex2 triplets, such as the diverse chaperones and chaperonins, peptidases, and proteases. Other proteins whose genes contain Ex1 and Ex2 triplets are involved in DNA replication, DNA processing, such as transcriptional factors, repressors or regulators, diverse DNA polymerases and DNA repairing enzymes. Signalling enzymes, such as kinases and phosphatases, and diverse receptors are also encoded by EGCs. Enzymes necessary for motility or cell division are encoded, importantly, by Ex1 and Ex2 triplets. Proteins that confer resistance against environmental factors or that currently are involved in pathogenicity or defence against other organisms, contain some portions encoded by Ex1 and Ex2 triplets. Finally, there exist poorly characterised proteins that being part of the EPh, some of them are named by the molecule they bind or process in which they participate, but there are also several hypothetical proteins (data not included, due to vagueness), DUFs and TIGR families that, albeit they are not fully characterised they must be certainly important given that portions of them existed previous to LUCA. Proteins whose portions are encoded by triplets of type Ex1 or Ex2 conform an Extended Proteome that comprises several of the proteins previously reported as constituents of the cellular repertoire of LUCA ([Mushegian, 1999](#); [Koonin, 2003](#); [Glansdorff et al., 2008](#); [Forterre et al., 2005](#); [Delaye et al., 2005](#); [Ouzounis et al., 2006](#)), some of them even vital for minimal synthetic organisms ([Pennisi, 2010](#)). We did recover fragments of the entire set of proteins that eventually would lead to the formation of LUCA an organism based on Wood-Ljungdahl or reductive Acetyl-CoA pathway, together with fragments of oxidative-phosphorylation chain proteins, DNA as information storage, to which reverse gyrase is essential, and the whole translation apparatus and several RNA modification enzymes, as well as ways to communicate with the environment, interchange molecules and signalling for the necessary changes based on environmental conditions ([Table 2](#)). Extended phenotypes encompass portions of proteins from proto-tRNAs, domains from ribosomal proteins, and proteins that eventually would shape the physiology of LUCA.

#### 3.2.1. Supplementary materials

**S1, S2, and S3** correspond to the full lists of proteins encoded to some extent by RNY, Ex1 and Ex2 triplets, respectively.

#### 3.3. Examples of the proteomes before LUCA

By focusing only on the names of proteins, it appears an almost complete overlapping among those encoded by triplets of each extended genetic code (EGC); however, the fragments retrieved from Ex1 are more but they are shorter, whereas fewer fragments were retrieved from the Ex2 but they are longer. Additionally, each EGC encodes slightly different or very different portions of the same protein, from which we can predict its structure and plausible ligands. Thereby, we captured different stages of enzymatic accretion, starting with generic RNY-encoded CSBS that eventually evolved towards the extended phenotype (EPh) encoded by Ex1 and Ex2 triplets.

Numerous ribosomal proteins contain portions encoded by EGCs, and some of them can be traced back even to RNY codification, as is the case of the ribosomal protein S11 shown in [Fig. 1](#).

It has been ascertained that several types of proteins originated from ribosomal proteins ([Lupas and Alva, 2017](#)); we found that some of those enzymes have portions encoded by both EGCs, and some of them even have portions identifiable as encoded by RNY triplets. We detected such evolution by enzymatic accretion, starting with the portion encoded by RNY triplets, which evolved towards a portion encoded by Ex1 triplets, or encoded by Ex2 triplets. Two of these proteins are triose-phosphate isomerase (TPI, [Fig. 2](#)) and Threonine-tRNA ligase (ThrRL, [Fig. 3](#)).

A similar exercise was carried out with the evolution of Threonine-tRNA ligase (ThrRL) ([Fig. 3](#)), that has also been proposed that evolved from a ribosomal protein ([Lupas and Alva, 2017](#)). Note the previous

**Table 2**

A non-exhaustive list of proteins with EGCs-encoded regions.

All subunits of several types of DNA topoisomerases and DNA gyrases
All subunits of several types of DNA gyrases
All subunits of many types of DNA polymerases (I, II, III, IV); DNA-dependant and RNA-dependant)
Enzymes for enabling competence and conjugation
Proteins of phage integration
DNA and RNA helicases
All the 20 tRNA ligases (aaRS)
All ribosomal proteins (of both 50S and 30S subunits)
Several rRNA modification enzymes (16S rRNA and 23S rRNA methyltransferases, and non-specific methyltransferases)
All factors of translation apparatus
Some metabolic proteins of different types of energetic metabolisms (carbon fixation, sulphur, methane, and nitrogen metabolism), including all subunits of enzymes of Wood-Ljungdahl pathway, and ATPases subunits.
All enzymes of glycolysis and TCA cycle
Several enzymes of amino acids metabolism
Several enzymes of nucleotides metabolism (for salvage, and <i>de novo</i> synthesis)
Several enzymes of lipids metabolism (for synthesis of lipid-chains of varied length)
Several types of peptidases and proteases
All chaperones and chaperonins (GroEL, GroES, DnaJ, DnaK)
Several enzymes for synthesis and modification of cell wall proteins (synthesis of peptidoglycan)
Several enzymes for synthesis and modification of membrane proteins (such as glycosyltransferases)
All components of ABC transporters for several types of molecules (ions, aa, sugars)
All components of PTS of several sugars
Varied antiporters and symporters
Mechanosensitive channels or ionic channels
Porins and aquaporins
Transporters for extrusion of nocive compounds
Kinases (Ser/Thr, Asp, His)
Phosphatases (Ser/Thr, Asp, His)
Proteins for flagellar movement
All proteins for cell division
Multiple functions / unknown functions
DUFs and proteins TIGR.
Hypothetical proteins

presented portion encoded by RNY triplets (**3A**), evolving towards the portion encoded by Ex1 triplets (**3B**), or encoded by Ex2 triplets (**3C**).

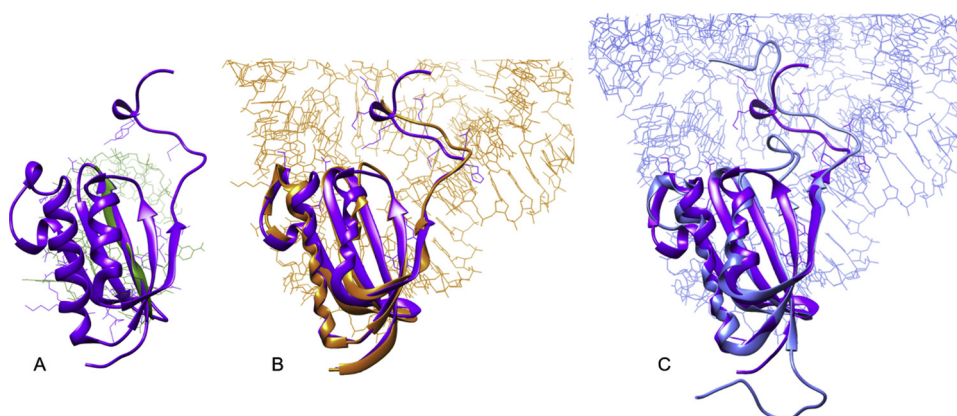
We also found that most of the proteins of the Wood-Ljungdahl methanogenic pathway are encoded by EGCs. The reductive acetyl-coenzyme A (AcCoA) metabolism was stated by Weiss et al. (2016) as the metabolism of LUCA although they did not find essential parts of its metabolic pathway. We found that AcCoA synthase (AcCoAs), a key enzyme in Wood-Ljungdahl pathway, has indeed portions encoded by Ex1 and Ex2 triplets, and therefore we reconstructed its structure and probable ligands. AcCoAs can be monomeric, dimeric, or tetrameric, depending on the organism. We found, for instance, that AcCoAs from *Deinococcus radiodurans* (Derad) is miss-assigned because it is more similar to Acetyl-CoA synthetase that it is not involved in Wood-Ljungdahl

pathway. On the other hand, *Methanocaldococcus jannaschii* (Mejan) has the only enzyme labelled as “acetyl-coenzyme A synthase” in our set of 26 organism; however, Mejan and *Methanosarcina acetivorans* (Macet) has many copies of proteins assigned as AcCoAs subunit alpha to delta, that contain portions encoded by RNY and both EGC. According with Lindahl and Chang (2001), subunit alpha reflects more accurately the phylogeny and functional diversity of AcCoAs, therefore we show the structure of the consensus sequence of alpha subunit of AcCoAs, encoded by each EGG in Fig. 4A and B.

Additionally, Wächtershäuser (1990) and Huber and Wächtershäuser (1997) proposed a primeval chemoautotrophic hot metabolism based on reduction of acetyl-coenzyme A in presence of FeS complexes, that could function as a primeval Wood-Ljungdahl pathway (Lindahl and Chang, 2001). Subunit gamma of modern AcCoAs binds the FeS compounds (2Fe<sub>2</sub>S, 4Fe<sub>4</sub>S, 4FeS, 2FeS, FeS, etc) which are necessary for metabolic reactions; while gamma subunit of Mejan contains a tiny portion encoded by RNY triplets (data not shown), Ex1 and Ex2 encode significant portions of gamma subunits from where a consensus was obtained and its structure predicted, as shown in Fig. 4C and D.

A crucial enzyme that can strongly indicate a thermophilic life style is reverse gyrase (McInerney, 2016), whose sequence we detected as partially encoded since the primeval genetic code (PGC), illustrated in Fig. 5A, and after, by both extended genetic codes (EGCs), shown in Fig. 5B and C. The finding becomes the more valuable in our dataset, because it serves to prove that our results are not artefactual in the sense that the same enzyme have portions encoded by both EGCs, but it shows how different the protein product of each code can be. It reveals that each EGC are complement one to the other not only in triplets or their corresponding amino acids but further in the peptides that each one encodes. Ex1 triplets encodes only tiny portions of the reverse gyrase of several organisms, the consensus sequence barely coincides to such enzyme, and the structure cannot be distinguished from randomness, as indicated in captions of Fig. 5B. In contrast, Ex2 triplets encodes a protein highly similar to the modern reverse gyrase, both in sequence and structure, as indicated in caption of Fig. 5C.

There is an apparent paradox regarding the amino acids (aa) encoded by each EGC and the aaRLs retrieved by those EGCs. For instance, one triplet of Ex1 type encodes the amino acid tryptophan (Trp) and it is not encoded by any Ex2 triplet (José et al., 2009), although tryptophan-tRNA ligase (TrpRL) appears as encoded by both EGCs. However, the structural reconstructions of the consensus sequences of TrpRL from each EGC, shown in Fig. 6, reveals that the protein encoded by Ex1 triplets is more reliable than the protein encoded by Ex2 triplets, that is reflected in thrice C-score for Ex1-TrpRL when compared to Ex2-TrpRL and better TM-score of Ex1-TrpRL over Ex2-TrpRL with respect to the current enzyme.



**Fig. 1.** In (A), RNY-encoded portion of S11 (green, C = 0.24, ligand: nucleotide) superimposed with the current protein S11 (4v90 purple; TM = 0.35501); in (B) Ex1-encoded portion of S11 (orange, C = 0.47, ligand: nucleotide) superimposed with the current protein S11 (4v90 purple; TM = 0.86730). In (C) Ex2-encoded portion of S11 (blue, C = 0.48, ligand: nucleotide) superimposed with the current protein S11 (4v90 purple; TM = 0.73278).



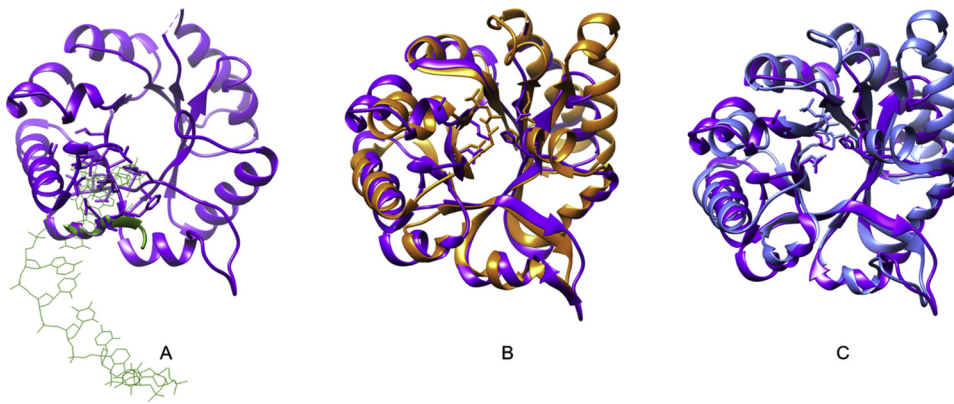


Fig. 2. In (A), RNY-encoded portion of TPI (green C = 0.56, ligand: nucleotide) superimposed with the current protein TPI (2h6r purple; TM = 0.49609); in (B), Ex1-encoded portion of TPI (orange C = 1.00, ligand: 3 PG) superimposed with the current protein TPI (2h6r purple; TM = 0.76436); in (C), Ex2-encoded portion of TPI (blue C = 0.99, ligand: 3 PG) superimposed with the current protein TPI (2h6r purple; TM = 0.76227).

#### 4. Discussion

We have characterised the proteomes in two distinctive stages of life, since the so-called RNA world, as RNY-encoded Cofactor Stabilising Binding Sites, up to the dawn of LUCA with Ex1 and Ex2 genomes that encode Extended Phenotypes presumably pertaining to progenotes. Our list of organisms includes the organisms used to prove the mathematical scale invariance of evolution (José et al., 2009), plus several other organisms with a wide variety of lifestyles and phylogenetically unrelated. We include pathogenic bacteria (such as *Borrelia burgdorferi*), parasitic bacteria (such as *Mycoplasma genitalium*), some of the most characterised bacteria (such as *Escherichia coli* or *Bacillus subtilis*), as well as the free-living organism with the smallest genome (ca. *Pelagibacter ubique*); in our list there are also several extremophile organisms, such as the high-pressure tolerant bacterium *Shewanella piezotolerans* or the square halophile archaeum *Haloquadratum walsbyi*, the radiation-loving *Deinococcus radiodurans*, as well as hyperthermophile archaea and bacteria. The varied life regimes implied that many proteins could be found, for example, in 5, or 8, or 17 out of the 26 organisms; accordingly, we included in the extended phenotype those proteins appearing in, at least, two out of the 26 organisms analysed. Remarkably, the proteins that appear in absolutely all 13 bacteria and 13 archaea are **ribosomal proteins and aaRL**, contrasting with the RNY-encoded proteins on which there are few of this type of proteins even in individual organisms. It indicates that the core of translation apparatus evolved well before LUCA but shortly after the beginnings of the ribonucleoprotein (RNP) world.

Our results are consistent with independent methodologies that reconstructed the probable metabolic pathways of LUCA via methods based in homology (Delaye et al., 2005). Other works reconstructed the ancestor sequences for each type of tRNA with RNY anticodons (Farias et al., 2016), and their results are embedded in ours. By examining modern protein structures, Lupas and Alva (2017) traced back ancestral motifs still present in ribosomal proteins, such as triose-phosphate isomerase, reported in this work. All these works reinforce the concept that a large fraction of genes within a genome are ultimately related by descent to a small number of genes that arose early in our evolutionary

history (Maynard-Smith 1998).

There are approaches that not only collect common genes or proteins, but that look for the translation products from concatenated proto-tRNAs (Farias et al., 2016); or that try to reconstruct the microbial ecology of LUCA (Weiss et al., 2016); or that trace the evolution of ribosomal proteins that eventually formed modern proteins, according with its structural resemblance (Lupas and Alva, 2017). Overall, the proteins that we found have been reckoned as ancient by other independent works. Our work has been characterised by tracing the evolution of genotype, before the conformation of the current standard genetic code (SGC), looking at the evolution of phenotype; in this manner, we were able to capture the evolution of proteomes in two distinct stages of life, since the so-called RNA world, as RNY-encoded CSBS, up to the dawn of LUCA with Ex1 and Ex2 triplets that encode an Extended Phenotype (EPh). Although few proteins began as CSBSs, most of them have portions only recognised by the later Ex1 and/or Ex2 triplets, with no differences among organisms for the same EGC. One of the principal differences is given by the presence of all three stop codons that only Ex1 present. Such EGC encodes so short fragments on some proteins of organisms, that it becomes difficult to construct a distinguishable protein structure because those fragments rarely overlap; whereas Ex2 encodes sufficiently long fragments that it covers almost the whole length of the proteins in most of the organisms; the proteins reGyr, ThrRL and TrpRL exemplify these differences between Ex1 and Ex2. For instance, Ex2 triplets encode a more reliable reGyr while Ex1 triplets produce a peptide close to randomness. In other cases, the triplets of type Ex1 encode slightly better structures than Ex2 triplets, such is the case of ThrRL that charges an amino acid encoded by both types of EGCs (Thr); TrpRL, on the other hand, charges an amino acid that it is not encoded by any triplet of type Ex2 (Trp) and the structure encoded by Ex1 triplets is actually three-fold reliable than Ex2-encoded protein. Other types of proteins like TPI and S11, with a small but significant portion encoded by RNY triplets and that later are encoded almost completely by both EGC, indistinctly, throughout its whole length. These types of proteins were constituted after RNA world but at the time that EGCs emerged, they had already been formed, or because the amino acids that constitute those proteins belong to both

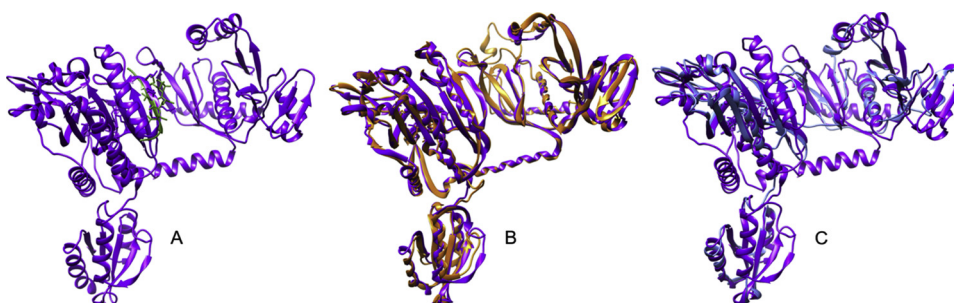
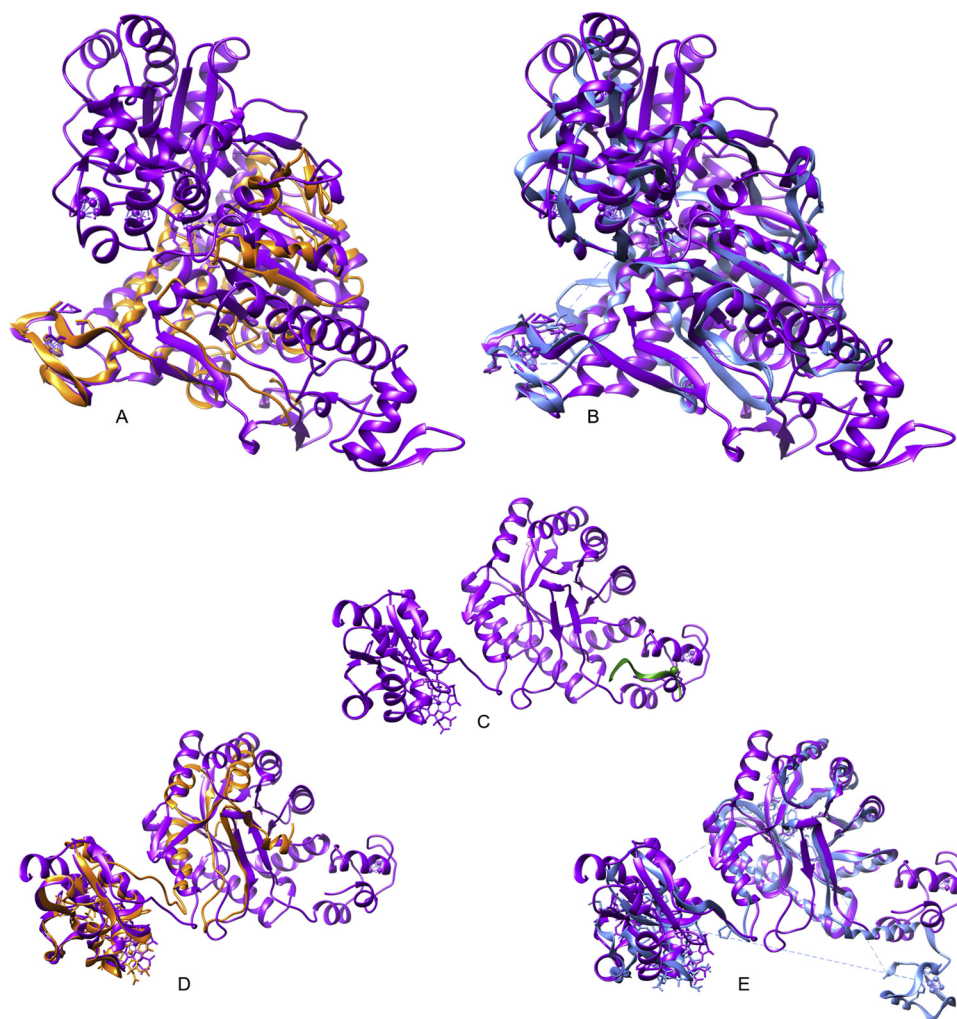


Fig. 3. In (A), RNY-encoded portion of ThrRL (green, C = 0.30, ligand: pept) superimposed with the current protein ThrRL (1qf6 purple; TM = 0.35501); ThrRL has different RNY-encoded portions in different organisms. Here, only the RNY encoded portion from *Thermococcus gammatolerans* EJ3 is depicted; in (B), Ex1-encoded portion of ThrRL (orange, C = 0.54, ligand: TSB) superimposed with current ThrRL (1qf6 purple, TM = 0.94090); in (C), Ex2-encoded portion of ThrRL (blue, C = 0.42, ligand: 409) superimposed with current ThrRL (1qf6 purple TM = 0.55094).



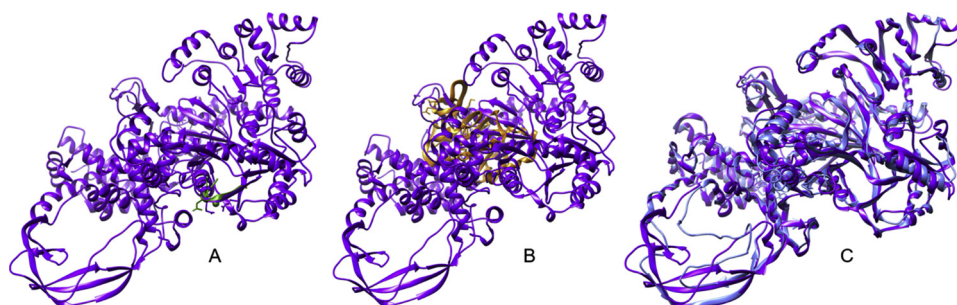
**Fig. 4.** A. In (A), Ex1-encoded portion of AcCoAs alpha subunit (orange,  $C = 0.22$ , ligand: SF4) superimposed with current AcCoAs (3cf4 purple,  $TM = 0.82924$ ); in (B), Ex1-encoded portion of AcCoAs alpha subunit (blue,  $C = 0.14$ , ligand: SF4) superimposed with current AcCoAs (3cf4 purple,  $TM = 0.84568$ ). In (C), RNY-encoded portion of AcCoAs gamma subunit (green,  $C = 0.33$ , ligand: His + Ca<sup>2+</sup>) superimposed with current AcCoAs (2ycl purple,  $TM = 0.35261$ ); in (D), Ex1-encoded portion of AcCoAs gamma subunit (orange,  $C = 0.54$ , ligand: mCOB) superimposed with current AcCoAs (2ycl purple,  $TM = 0.80340$ ); in (E), Ex2-encoded portion of AcCoAs gamma subunit (blue,  $C = 0.14$ , ligand: B12+SF4+HH2) superimposed with current AcCoAs (3cf4 purple,  $TM = 0.74329$ ).

types of EGCs (José et al., 2009). A third type of proteins like AcCoAs, composed of several subunits, each one with different portions encoded by each EGC. Such portions have similar TM-score but each type of triplets (Ex1 or Ex2) encode slightly different sections on the same subunit. AcCoAs is the key enzyme of Wood-Ljungdahl pathway, that was not found by Weiss et al. (2016) despite that reductive acetyl-CoA metabolism has been proposed as primordial (Lindahl and Chang, 2001).

As we described with examples and can be seen in Supplementary material S1, the proteins with portions encoded by RNY triplets are varied. We obtained only one short fragment per protein and the majority are non-overlapping among them, yet such Ur-proteome corresponds to the primitive bindome. CSBSs bind ancestral and prebiotic molecules, such as those with nucleotide moieties (Jadhav and Yarus, 2002; Sharov, 2016), or small peptides (Tamura and Schimmel, 2003;

Wieczorek et al., 2017) or generic nucleotide chains (Agmon, 2016) randomly generated, even metallic ions (Fedor, 2002), water or small organic molecules (Barton et al., 2007; Powney and Sutherland, 2011). Primitive proteins were not directly involved in catalysis, but they provided stable conformations (Shibue et al., 2018). During the early stage of evolution, amino acids were not selected for their ability to promote catalytic reactions, but for allowing the formation of stable and soluble tertiary structures (Doig, 2017; Shibue et al., 2018). This notion implies that other molecules such as RNA, cofactors and metals (Bray et al., 2019), might have played a central role in catalytic function. Most of the amino acids that existed in the Ur-proteome did not contain functional side chains that are important for catalysis (Doig, 2017).

According to our results, the proteomes before LUCA were wide and varied, yet incomplete. As shown in Table 2 and Supplementary



**Fig. 5.** In (A), RNY-encoded portion of reGyr (green,  $C = 0.49$ , ligand: Ca<sup>2+</sup>) superimposed with the current protein TPI (4ddt purple;  $TM = 0.29435$ ); in (B), Ex1-encoded portion of reGyr (orange,  $C = 0.07$ , ligand: Mg<sup>2+</sup>) superimposed with the current protein TPI (4ddt purple;  $TM = 0.27493$ ); in (C), Ex2-encoded portion of reGyr (blue,  $C = 0.15$ , ligand: TMP) superimposed with the current protein TPI (4ddt purple;  $TM = 0.96115$ ).

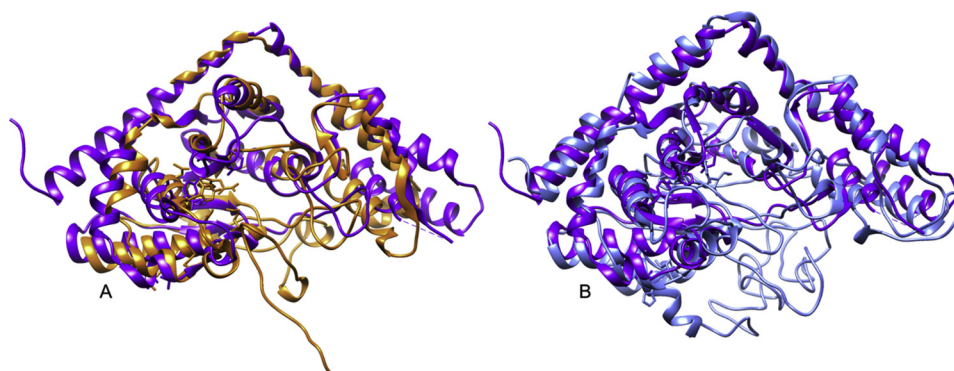


Fig. 6. A. In (A), Ex1-encoded portion of TrpRL (orange, C = 0.68, ligand: Trp) superimposed with current TrpRL (2e17 purple, TM = 0.70574); in (B), Ex2-encoded portion of TrpRL (blue, C = 0.22, ligand: Trp) superimposed with the current TrpRL (2e17 purple, TM = 0.57325).

materials S2 and S3, it seems that LUCA and previous organisms were already complex organisms. For instance, there appears an almost complete repertoire of translation process proteins, including ribosomal proteins, RNA modification enzymes, and amino acyl-tRNA synthetases. There appear several types of communication passages such as antiporters, symporters, channels, and pores. There exist some proteins involved in synthesis of membrane and cell wall, although bacteria and archaea share few enzymes of this category because of the different nature of its membranes. Proteins involved in signal transduction have portions encoded by EGCs, particularly several types of kinases and many phosphatases, as well as some metabolites-regulated proteins. There are also proteins involved in re-folding and degrading peptides. Some proteins that confer resistance to the organism against environmental stress, or that confer movement to the organisms by using flagella, appear also as part of the Eph. Other proteins partially encoded by EGCs are the diverse highly conserved hypothetical proteins, because its degree of conservation goes well before the appearance of LUCA, and because they contain portions encoded by Ex1 and Ex2 triplets. On the other side, it seems paradoxical that proteins of DNA processing are encoded by EGCs, because LUCA is, by definition, an organism on which DNA is the inheritance molecule that follows the current SGC (Woese, 1998), while Ex1 and Ex2 triplets evolved, by definition, before LUCA (José MV et al. 2009). To our knowledge, a possible explanation comes from the hypothesis of “Primordial Virus World Scenario”, proposed by Forterre (2005, 2006, 2010) and supported by Koonin, 2009, Zimmer (2006) and others (Durzyńska and Goździcka-Józefiak, 2015). Such proposal indicates that DNA arose first in viruses as a molecular mutation of RNA to evade the “immune” systems of the proto-organisms still dependant on RNA as principal molecule; in this manner, viruses have shaped the major evolutionary transitions since the earliest life forms (Forterre, 2006). This viral scenario is supported by the fact that several proteins of CRISPR-Cas system, considered the “immune” or defense system of prokaryotes (Makarova et al., 2013; Marraffini, 2015), have portions encoded by both extended genetic codes (Supplementary materials S2 and S3). Thus, viral or mobile elements evolved at the same time as cellular organisms, that eventually invaded and that could be one of the reasons why some phage proteins or phage-related proteins, integrated on genomes, are encoded also by both EGCs. The finding of RNA-dependent RNA polymerases since the initial stages of the origins of life (Farias et al., 2017), and our finding of the defense system provided by a primitive CRISPR-Cas complex, and some other viral proteins, open up new avenues of research regarding the ecological interactions of capsid-encoding (virus) and ribosome encoding (cells) entities, that certainly shaped their evolution.

Important works on the gradual transformation of RNA into DNA support also our findings, such as mutated metabolic enzymes guiding intermediate steps in the transition from RNA to DNA. Some of the most important molecules for this transition would be: RNA methylases,

ribonucleotide reductase (RNR), DNA/RNA polymerases that share common catalytic mechanism, DNA/RNA helicases that belong to the same superfamily (Poole et al., 2000, 2002; Poole and Logan, 2005; Poole et al., 2014). Actually “much of the DNA synthesis machinery may have been recruited from RNA synthesis” (Poole et al., 2000), and we found all those enzymes indicated as necessary for the RNA to DNA transition are encoded by Ex1 or Ex2 triplets. This evolutionary transition has been successfully simulated *in silico*, considering dynamic environmental scenarios (Ma et al., 2015). The presence of reverse gyrase (reGyr) that we found reliably encoded by Ex2 triplets, could be an additional support to disentangle the apparent paradox of having DNA processing enzymes in our dataset. It turns out that reGyr, is involved in packaging DNA by hyperthermophile organisms (Lulchev and Klostermeier, 2014), or to renature DNA (Hsieh and Plank, 2006), but it is also essential for packaging the genome of the viral-like particle SSV1 of archaea *Sulfolobus* (Nadal et al., 1986); The latter could signal its original role, marking out the approximate time of RNA to DNA transition, and reinforcing the proposition that viruses guided evolution in such extremophile environmental conditions that has been appointed as a plausible origin of life (Rice et al., 2001).

The foregoing observations, refer to the nominal appearing of proteins, however, we retrieved the specific portions of each protein, encoded on each evolutionary stage, and even if the proteins were not complete before LUCA, the recovered fragments have sufficient sequence identity to be recognised as one protein or another, although the structural reconstructions of consensus sequences are not accurate in all cases. Furthermore, as we previously indicated, the retrieved LUCA proteins from different approaches, are embedded in our extended phenotype (Eph). For instance, all the set of proteins that were encoded by proto-tRNAs (Farias et al., 2016), except “ornithine decarboxylase antizyme”, appear in our dataset (found in S2 and S3 that can be contrasted with Farias et al., 2016). Qualitatively speaking, some of the proteins encoded by triplets of the type Ex1 or Ex2 had been previously mentioned as components of the microbial ecology of LUCA (Weiss et al., 2016); in particular, we encountered fragments of most of the proteins of the Wood-Ljungdhal pathway, that was stated as the metabolism of LUCA despite that they were not found in the exhaustive comparison made by Weiss et al. (2016). Since we found fragments of all those proteins, we cannot argue that certain metabolic pathway must have existed, albeit it is evident that the raw materials were already available. What is evident is that would-be-enzymes already existed. Finally, the proteins that would have begun as ribosomal ones (Lupas and Alva, 2017), according to its structural resemblance, we did recover portions of such proteins, encoded by the EGCs.

In summary, we have delineated plausible evolutionary routes of the proteome from its beginnings to the dawn of LUca. First, nucleotide chains were randomly generated (Agmon, 2016), but when they began to interact with the prebiotically-produced amino acids (Miller, 1953), it emerged the RNY as the primeval genetic code (PGC), which gave

origin to CSBSs as the PPH on one side, whereas some of those RNA chains were constituted on the other side as proto-tRNAs that were translated (Farias et al., 2016). In this context, we also place the “omnipresent motifs” that Sobolevsky et al. (2013) described as the proteome of LUCA, as we found such peptides can be encoded by RNY triplets (Palacios-Pérez et al., 2018) as CSBSs that arose earlier than the emergence of LUca. All those peptide sets could have pertained to distinct types of progenotes (Woese, 1998) and then to an acellular entity named First Universal Common Ancestor (FUCA), that had a peptidyl transferase center that started complex translational apparatus (Farias et al., 2014) and therefore could synthesise peptides, and then proteins, leveraging the environmental feedback by means of bare transporters and elementary signalling networks (Prosdocimi et al., 2018).

## 5. Conclusions

We have portrayed the step-wise evolution of proteomes towards the very dawn of a LUCA richer and more complex than previously considered, and then with more potential to speciate towards not so plain life forms. We illustrated with some examples (S11 ribosomal protein, triose-phosphate isomerase, ThrRL, TrpRL, reverse gyrase, and AcCoA synthase) the validity of our approach, by elucidating their three-dimensional structure, their ligands (high C-scores), and with the comparisons with the modern crystallographic structures of the corresponding molecules. These proteins were already functional and structurally very similar (high TM-scores) to their modern versions. The first proteins to be consolidated were the ribosomal ones, that initially evolved as chaperones to give structural stability to ribonucleic chains (Lupas and Alva, 2017); those proteins served then as raw material to generate some other peptides. At the same time, PGC evolved and mutated through two pathways, generating both EGCs that encoded proteins in an organism more complex than previously thought. Such incomplete proteins at the dawn of LUCA, gradually built the set of proteins that LUCA used, already based on Wood-Ljungdahl metabolic pathway (Weiss et al., 2016), with a complex cell wall, using DNA as information storage and RNA only as translational molecule, and that respond to environmental stimuli. The recovered portions of many types of proteins encoded in each evolutionary stage may well pertain to FUCA, be the PGC made of RNY triplets that encode CSBSs, or the two intermediate EGCs, previous to the formation of the SGC that LUCA already had, and whose repertoire of proteins implied combinations and growing of the molecules already consolidated in previous evolutionary steps. Finally, to the best of our knowledge, we are closer to seal the hitherto existing gap between the pre-biotal and proto-biotal stages, tracing the evolution of phenotype based on the evolution of genotype, beginning with the PGC and then with two intermediate EGC which finally shaped the current SGC.

## Author's contributions

MPP and MVJ. designed the study and analyses. MPP performed the analyses. MPP and MVJ wrote the manuscript.

## Competing interests

The authors declare no competing financial interests.

## Acknowledgements

MPP is a doctoral student from Programa de Doctorado en Ciencias Biomédicas (PDCB), Universidad Nacional Autónoma de México (UNAM) with support of the fellowship 694877 from CONACYT. MVJ was financially supported by PAPIIT-IN201019. We thank M.S. Fernando Andrade-Díaz and M.I. Juan R. Bobadilla for computer assistance. MPP also thanks to Sohan Jheeta by his financial support for

poster presentation at ISSOL 2017, that was the point of departure of the present work. Authors thank to Arturo Becerra and Luis Eguiarte for their feedback during this project, and to the Theoretical Biology Group for helpful discussions.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.biosystems.2019.04.007>.

## References

- Agmon, I., 2016. Could a proto-ribosome emerge spontaneously in the prebiotic world? *Molecules* 21, e1701.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Barton, N.H., Briggs, D.E.G., Eisen, J.A., Goldstein, D.B., Patel, N.H., 2007. *Evolution*. Cold Spring Harbor Laboratory Press, USA.
- Bray, M.S., Timothy, K., Lenz, T.K., Haynes, J.W., Bowman, J.C., Petrov, A.S., Reddi, A.R., Hud, N., Williams, L.D., Glass, J.B., 2019. Multiple prebiotic metals mediate translation. *Proc. Natl. Acad. Sci. U. S. A.* 115, 12164–12169. <https://doi.org/10.1073/pnas.1803636115>.
- Carter Jr, C.W., 2014. Urzymology: experimental access to a key transition in the appearance of enzymes. *J. Biol. Chem.* 289, 30213–30220.
- Cech, T.R., 2012. The RNA worlds in context. *Cold Spring Harb. Perspect. Biol.* 4, a006742.
- Delaye, L., Becerra, A., Lazzano, A., 2005. The last common ancestor: what's in a name? *Orig. Life Evol. Biosph.* 35, 537–554.
- Doig, A.J., 2017. Frozen, but no accident - why the 20 standard amino acids were selected. *FEBS J.* 284, 1296–1305. <https://doi.org/10.1111/febs.13982>.
- Durzyńska, J., Goździcka-Józefiak, A., 2015. Viruses and cells intertwined since the dawn of evolution. *Virol. J.* 12, 169.
- Eigen, M., Lindemann, B., Winkler-Oswatitsch, Clarke, R.H.C., 1985. Pattern analysis of 5S rRNA. *Proc. Natl. Acad. Sci. USA* 82, 2437–2441. <https://doi.org/10.1073/pnas.82.8.2437>.
- Eigen, M., Schuster, P., 1978. The hypercycle, A principle of natural self-organization, Part C: the realistic hypercycle. *Naturwissenschaften* 65, 341–369.
- Farias, S.T., Régo, T., José, M.V., 2014. Origin and evolution of the Peptidyl Transferase Center from proto-tRNAs. *FEBS Open Bio* 4, 175–178.
- Farias, S.T., Régo, T.G., José, M.V., 2016. tRNA core hypothesis for the transition from the RNA world to the ribonucleoprotein world. *Life Basel (Basel)* 6, e15.
- Farias, S.T., Dos Santos Jr., A., Régo, T.G., José, M.V., 2017. Origin and evolution of RNA-dependent RNA polymerase. *Frontiers in Genetics* 28 (125). <https://doi.org/10.3389/fgene.2017.00125>. ISSN: 1664-802.1.
- Fedor, M.J., 2002. The role of metal ions < AT > /AT > in RNA catalysis. *Curr. Opin. Struct. Biol.* 12, 289–295.
- Forterre, P., 2005. The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. *Biochimie* 87, 793–803.
- Forterre, P., 2006. The origin of viruses and their possible role in major evolutionary transition. *Virus Res.* 117, 5–16.
- Forterre, P., 2010. Defining life: the virus viewpoint. *Orig. Life Evol. Biosph.* 40, 151–160.
- Forterre, P., Gribaldo, S., Brochier, C., 2005. Luca: the last universal common ancestor. *Med. Sci. (Paris)* 21, 860–865.
- Glandsdorff, N., Xu, Y., Labedan, B., 2008. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol. Direct* 3, 29.
- Hsieh, T.S., Plank, J.L., 2006. Reverse gyrase functions as a DNA renaturase: annealing of complementary single-stranded circles and positive supercoiling of a bubble substrate. *J. Biol. Chem.* 281, 5640–5647.
- Huber, C., Wächtershäuser, G., 1997. Activated acetic acid by carbon fixation on (Fe,Ni)S under primordial conditions. *Science* 276, 245–247.
- Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., Jensen, L.J., von Mering, C., Bork, P., 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucl. Acids Res.* 44, D286–D293.
- Jadhav, V.R., Yarus, M., 2002. Coenzymes as coribozymes. *Biochimie* 84, 877–888.
- Johnson, A.P., Cleaves, H.J., Dworkin, J.P., Glavin, D.P., Lazzano, A., Bada, J.L., 2008. The Miller volcanic spark discharge experiment. *Science* 322, 404.
- José, M.V., Morgado, E.R., Govezensky, T., 2007. An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull. Math. Biol.* 69, 215–243.
- José, M.V., Govezensky, T., García, J.A., Bobadilla, J.R., 2009. On the evolution of the standard genetic code: vestiges of critical scale invariance from the RNA world in current prokaryote genomes. *PLoS One* 4, e4340.
- José, M.V., Morgado, E.R., Govezensky, T., 2011. Genetic hotels for the standard genetic code: evolutionary analysis based upon novel three-dimensional algebraic models. *Bull. Math. Biol.* 73, 1443–1476.
- José, M.V., Zamudio, G.S., Palacios-Pérez, M., Bobadilla, J.R., de Farias, S.T., 2015. Symmetrical and thermodynamic properties of phenotypic graphs of amino acids encoded by the primeval RNY code. *Orig. Life Evol. Biosph.* 45, 77–83.
- José, M.V., Zamudio, G.S., Morgado, E.R., 2017. A unified model of the standard genetic code. *Ro. Soc. Open Sci.* 4 <https://doi.org/10.1098/rsos.160908>. 160908.

- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. *Nucl. Acids Res.* 34, D354–D357.
- Koonin, E.V., 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* 1, 127–136.
- Koonin, E.V., 2009. On the origin of cells and viruses: primordial Virus World scenario. *Ann. N. Y. Acad. Sci.* 1178, 47–64.
- Lehmann, J., 2002. Amplification of the sequences displaying the pattern RNY in the RNA world: the translation & translation/replication hypothesis. *J. Theor. Biol.* 219, 521–537.
- Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Mi Park, Y., iBUSE, N., Lopez, R., 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* 43, W580–W584.
- Lindahl, P.A., Chang, B., 2001. The evolution of acetyl-CoA synthase. *Orig. Life Evol. Biosph.* 31, 403–434.
- Lulchev, P., Klostermeier, D., 2014. Reverse gyrase—recent advances and current mechanistic understanding of positive DNA supercoiling. *Nucleic Acids Res.* 42, 8200–8213.
- Lupas, A.N., Alva, V., 2017. Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *J. Struct. Biol.* 198, 74–81.
- Ma, W., Yu, C., Zhang, W., Wu, S., Feng, Y., 2015. The emergence of DNA in the RNA world: an in-silico simulation study of genetic takeover. *BMC Evol. Biol.* 15, 272.
- Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2013. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* 41, 4360–4377.
- Marraffini, L.A., 2015. CRISPR-Cas immunity in prokaryotes. *Nature* 256, 55–61.
- McInerney, J.O., 2016. A four-billion-year-old metabolism. *Nature Microbiol.* 1, 16139.
- Merkel, R., Sterner, R., 2016. Ancestral protein reconstruction, techniques and applications. *Biol. Chem.* 397, 1–21.
- Miller, S.L., 1953. A production of amino acids under possible primitive earth conditions. *Science* 117, 528–529.
- Moretti, S., Armougom, F., Wallace, I.M., Higgins, D.G., Jongeneel, C.V., Notredame, C., 2007. The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Res.* 35, W645–W648.
- Mushegian, A., 1999. The minimal genome concept. *Curr. Opin. Genet. Dev.* 9, 709–714.
- Nadal, M., Mirambeau, G., Forterre, P., Reiter, W.D., Duguet, M., 1986. Positively supercoiled DNA in a virus-like particle of an archaeobacterium. *Nature* 321, 256–258.
- Okonechnikov, K., Golosova, O., Fursov, M., 2012. The UGENE team. *Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics* 28, 1166–1167. <https://doi.org/10.1093/bioinformatics/bts091>.
- Ouzounis, C.A., Kunin, V., Darzentas, N., Goldovsky, L., 2006. A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective. *Res. Microbiol.* 157, 57–68.
- Pagel, M.D., Pomiankowski, A., 2008. *Evolutionary Genomics and Proteomics*. Sinauer Associates, Sunderland, MA, USA.
- Palacios-Pérez, M., Andrade-Díaz, F., José, M.V., 2018. A Proposal of the Ur-proteome. *Orig. Life Evol. Biosph.* 48, 245–258. <https://doi.org/10.1007/s11084-017-9553-2>. Epub 2017 Nov 10.
- Pennisi, E., 2010. Synthetic genome brings new life to bacterium. *Science* 328, 958–959.
- Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612.
- Poole, A., Logan, D.T., 2005. Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA genome? *Mol. Biol. Evol.* 22, 1444–1455.
- Poole, A., Penny, D., Sjöberg, B., 2000. Methyl-RNA: an evolutionary bridge between RNA and DNA? *Chem. Biol.* 7, R207–R216.
- Poole, A., Logan, D.T., Sjöberg, B., 2002. The evolution of the ribonucleotide reductases: much ado about oxygen. *J. Mol. Evol.* 55, 180–196.
- Poole, A.M., Horinouchi, N., Catchpole, R.J., Si, D., Hibi, M., Tanaka, K., Ogawa, J., 2014. The case for an early biological origin of DNA. *J. Mol. Evol.* 79, 204–212.
- Powner, M.W., Sutherland, J.D., 2011. Prebiotic chemistry: a new modus operandi. *Philos. Trans. R. Soc. Lond. B* 366, 2870–2877.
- Prosdocimi, F., José, M.V., Farias, S.T.D., 2018. Be Introduced to the First Universal Common Ancestor (FUCA): The Great-Grandmother of LUCA (Last Universal Common Ancestor). Preprints <https://doi.org/10.20944/preprints201806.0035.v1>. 2018060035.
- Rice, G., Stedman, K., Snyder, J., Wiedenheft, B., Willits, D., Brumfield, S., McDermott, T., Young, M.J., 2001. Viruses from extreme thermal environments. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13341–13345.
- Rodin, S.N., Ohno, S., 1997. Four primordial modes of tRNA synthetase recognition, determined by the (G,C) operational code. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5183–5188. <https://doi.org/10.1073/pnas.94.10.5183>.
- Sharov, A.A., 2009. Coenzyme autocatalytic network on the surface of oil microspheres as a model for the origin of life. *Int. J. Mol. Sci.* 10, 1838–1852.
- Sharov, A.A., 2016. Coenzyme world model of the origin of life. *Biosystems* 144, 8–17.
- Shepherd, J.C., 1981a. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. *J. Mol. Evol.* 17, 94–102.
- Shepherd, J.C., 1981b. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci. U. S. A.* 78, 1596–1600.
- Shibue, R., Sasamoto, T., Shimada, M., Zhang, B., Yamagishi, A., Akanuma, S., 2018. Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep.* 8, 1227. <https://doi.org/10.1038/s41598-018-19561-1>.
- Sobolevsky, Y., Guimarães, R.C., Trifonov, E.N., 2013. Towards functional repertoire of the earliest proteins. *J. Biomol. Struct. Dyn.* 31, 1293–1300.
- Szathmáry, E., 2007. Coevolution of metabolic networks and membranes: the scenario of progressive sequestration. *Phil. Trans. R. Soc. B* 362, 1781–1787.
- Tamura, K., Schimmel, P., 2003. Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptors. *PNAS* 100, 8666–8669.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41.
- Thornton, J.W., 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.* 5, 66–375.
- Trifonov, E., 2004. The triplet code from first principles. *J. Biomol. Struct. Dyn.* 22, 1–11.
- Wächtershäuser, G., 1988. Before enzymes and templates: theory of surface metabolism. *Microb. Rev.* 52, 452–484.
- Wächtershäuser, G., 1990. Evolution of the first metabolic cycles. *Proc. Natl. Acad. Sci. U. S. A.* 87, 200–204.
- Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., Martin, W.F., 2016. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* 1, 16116. <https://doi.org/10.1038/nmicrobiol.2016.116>.
- Wieczorek, R., Adamala, K., Gasperi, T., Polticelli, F., Stano, P., 2017. Small and random peptides: an unexplored reservoir of potentially functional primitive organocatalysts. The case of seryl-histidine. *Life Basel (Basel)* 7, E19.
- Woese, C., 1998. The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6854–6859.
- Xu, D., Zhang, Y., 2011. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys. J.* 101, 2525–2534.
- Yang, J., Roy, A., Zhang, Y., 2013a. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 29, 2588–2595.
- Yang, J., Roy, A., Zhang, Y., 2013b. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* 41, D1096–D1103.
- Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9, 40.
- Zhang, Y., Skolnick, J., 2005. TM-align: a protein structure alignment algorithm based on TM-score. *Nucleic Acids Res.* 33, 2302–2309.
- Zimmer, C., 2006. Did DNA come from viruses? *Science* 312, 870–872.

## 5. DISCUSIÓN GENERAL

Teóricamente se ha demostrado que el código genético puede evolucionar desde el primordial RNY (Eigen and Schuster, 1978) hasta el SGC mediante los dos ExGCs, ello debido a que las propiedades estadísticas de uno a otro se preservan; adicionalmente, la evolución se deduce de la ruptura de simetrías en los modelos algebraicos tridimensionales de los codones (José et al., 2009; José et al., 2011).

Los modelos algebraicos se pueden también elaborar utilizando los aa codificados por los tripletes de algún tipo de código genético, en lugar de los codones correspondientes. De esta manera, se elaboraron grafos fenotípicos utilizando los 8 aa codificados por los 16 tripletes del tipo RNY; dichos aa son pequeños y la mayoría pudieron ser sintetizados prebióticamente; además, aunque poseen propiedades fisicoquímicas diversas, los grupos de simetría de los 12 ordenamientos posibles corresponden sólo al requerimiento polar (PR, por las siglas en inglés de *polar requirement*) de los aa, que los agrupa de dos en dos. Al considerar las medidas de centralidad de los 12 ordenamientos mencionados, se observa que todos los grafos son altamente simétricos estructuralmente, por lo que se deduce que los 8 aa son igualmente relevantes en el proceso evolutivo y la eliminación de tan solo uno tiene una fuerte selección en contra, lo que indica que el código RNY estaba *congelado*; aunque también se observa que el único aminoácido que muestra una centralidad atípica es Ala, con el menor valor de *closeness* –la distancia promedio de un vértice a los otros–, lo que puede indicar el punto de la ruptura de las simetrías de estos modelos algebraicos, lo que plausiblemente le permitió evolucionar (José et al., 2015), precisamente hacia los ExGCs.

El análisis de grafos mencionado permite entonces representar el fenotipo en aa y visualizar aquel a partir del que pudo evolucionar el código genético primitivo. Otra manera de investigar el fenotipo correspondiente al genotipo RNY es analizar las proteínas codificadas; dado que se sabe que un genoma completo posee las mismas propiedades estadísticas que un genoma que sólo contiene tripletes RNY, mientras ambos sean del mismo organismo, se pudieron obtener entonces las proteínas codificadas por ese genoma reducido (Palacios-Pérez et al., 2018).

Para reconstruir el **Ur-proteoma** se utilizó un único organismo —la bacteria *Streptococcus agalactiae* A909, especie que se escogió inicialmente para establecer el concepto de pan-genoma (Medini et al., 2005; Tettelin et al., 2005)—, a pesar de lo cual fue posible revelar el proteoma primigenio, hallazgos que se confirmaron en un posterior artículo (también propio). Cuando se comparó el genoma depurado con tripletes RNY versus el genoma original del organismo para obtener las proteínas codificadas por el primer código genético, no se encontró una sola proteína que en su totalidad fuese codificada por tripletes RNY —del **PGC**—; además, contrario a lo que inicialmente se consideraba a partir de comparaciones de la fisicoquímica de enzimas diversas (Jensen, 1976; Yčas, 1974), los sitios catalíticos activos de las proteínas no son tan antiguos como el primer código genético y sí lo son aquellos sitios de unión a diversas moléculas pequeñas primigenias.

Se ha determinado que los cofactores nucleotídicos —como NAD o NADP, FMN o FAD, ATP, o Coenzima A— pueden ser sintetizados prebióticamente y de hecho pudieron ser los primeros catalizadores (Szathmáry, 1990; Jadhav and Yarus, 2002; Sharov, 2009; Sharov, 2016) durante el mundo del RNA (Gilbert, 1986); mientras que los primeros aa, que pudieron haber sido sintetizados a partir de elementos fundamentales (Miller, 1953; Lazcano and Miller, 1999; Lazcano and Bada, 2003), habrían servido a su vez como cofactores de los oligonucleótidos (Szathmáry et al., 1999), dando lugar a ribonucleoproteínas (Di Giulio, 1997; Cech, 2009); cabe destacar que varios de esos aa son codificados por tripletes del tipo RNY (José et al., 2015). Por otro lado, pequeñas moléculas prebióticas como el glicerol, pudieron formar algunos de los primeros polímeros biológicos (Weber, 1989; Brisson et al., 2001) como ácidos grasos que a su vez formaron estructura membranosas (Segré et al., 2001), estabilizadas por los aa sintetizados prebióticamente (Cornell et al., 2019), donde se desarrollaban primitivos ciclos catalíticos impulsados por las coenzimas inicialmente mencionadas (Wächtershäuser, 1988b; Wächtershäuser, 1990). Por un lado se ha propuesto que los aa y nt prebióticos pudieron formar biopolímeros como oligopéptidos u oligonucleótidos sintetizados mediante ciclos húmedos-secos (*wet-dry*) (Campbell et al., 2019; Damer and Deamer, 2019) que, junto con los otros tipos de polímeros mencionados, generaron ciclos mutuamente catalíticos en los que diversos compuestos primigenios catalizaban la síntesis

de otros (Segré and Lancet, 2000; Segré et al., 2001; Lancet et al., 2018). Reacciones a las que pudo contribuir la catálisis mediada por iones metálicos (Fedor, 2002; Wächtershäuser, 2014; Bray et al., 2018). Ya sea que todos los componentes se encontrasen en la Tierra primitiva, o bien que algunos fuesen transportados por meteoritos que aleatoriamente impactaron contra el incipiente planeta (Urey, 1966; Cronin and Moore, 1971; Miller et al., 1976).

Todas las moléculas prebióticas mencionadas eran unidas y estabilizadas por el **Ur-proteoma** codificado por cadenas de ribonucleótidos que siguen un patrón del tipo RNY, por lo que este **bindoma** (préstamo de la palabra inglesa *bind*) resulta ser una colección de "Sitios de Unión y Estabilización de Cofactores" (**Cofactor Stabilising Binding Sites, CSBSs**) que conformados principalmente por los aa G, A, D, V, S, T, N e I; hallazgo que no solamente se basa en la localización del CSBS en la secuencia de las proteínas, sino en la reconstrucción estructural de esos péptidos y la predicción de las moléculas que pueden unir (Palacios-Pérez et al., 2018). Dichos CSBSs son incluso más cortos que los 30 aa que se ha predicho pudieron tener de longitud máxima los péptidos más antiguos (Trifonov, 2006).

Uno de los hallazgos más inesperados en la reconstrucción del **Ur-proteoma** (Palacios-Pérez et al., 2018) es que una porción de una metalo-beta-lactamasa (MβL) es codificada por tripletes RNY, aunque ese fragmento unía péptidos que pudieron polimerizarse también prebióticamente (Tamura and Schimmel, 2003); investigaciones sobre el origen de las βL indican un origen precámbrico (Risso et al., 2017) mientras que el primordio de tales proteínas, de acuerdo con nuestros resultados, se remonta hasta a los orígenes del código genético aunque con una función inicial distinta.

Anteriormente se llevó a cabo una comparación de secuencias de las proteínas de múltiples especies de bacterias y arqueas y se encontraron octámeros omnipresentes (Sobolevsky and Trifonov, 2006). Éstos octámeros fueron reconstruidos estructuralmente y las moléculas que unen fueron predichas, utilizando la metodología aplicada a los péptidos codificados por tripletes RNY y se encontró además que tales octámeros están compuestos principalmente por aa codificados por tripletes RNY (Palacios-Pérez et al., 2018).



Anteriormente se había planteado que en la estructura de enzimas modernas —de los metabolismos de purinas, de pirimidinas, de profirina, de clorofila y de carbohidratos—, están insertados los plegamientos proteicos más antiguos, que son dominios de unión a pequeños ligandos (Ji et al., 2007; Ma et al., 2008). Se ha planteado que ese tipo de dominios estaban presentes en el LUCA (Goldman et al., 2013), cuando todo indica que las estructuras capaces de unir ligandos prebióticos se formaron más bien cerca del origen de la vida (Chatterjee, 2016), ya que se ha mostrado que esos CSBSs pueden ser codificados por el PGC.

La lista de proteínas con porciones codificadas por tripletes RNY resultó ser una ecléctica colección de resultados, que implica que la estabilización de moléculas prebióticas apareció antes del uso catalítico de las mismas. Es necesario apuntar que aun cuando los fragmentos recuperados no sólo contienen los 8 aa codificados por RNY, sí están compuestos principalmente por esos aa en diversas combinaciones.

Para el trabajo de evolución del proteoma antes descrito, se utilizaron únicamente organismos denominados procariontes, con base en su morfología (Pace et al., 2012) *i.e.* arqueas y bacterias, ya que se ha señalado que los organismos con compartimentos internos rodeados por membranas y un núcleo —eucariontes— no están cerca de la raíz donde se ubicaría LUCA y son, de una u otra manera según enfoques distintos, un filo derivado (Di Giulio, 2011; Williams et al., 2012; Raymann et al., 2015; Spang et al., 2015; Spang and Ettema, 2016; Cornish-Bowden and Cárdenas, 2017; Imachi et al., 2020).

La colección de péptidos de los 26 organismos es similar a lo encontrado utilizando un solo genoma, es decir, el PGC codificaba péptidos pequeños capaces de unir pequeños ligandos que actualmente fungen como cofactores, dichos **CSBSs** están presentes en proteínas de diversos procesos metabólicos y en el caso de complejos multiméricos, sólo algunas subunidades tienen alguna porción codificada por tripletes RNY. Por otro lado, las proteínas codificadas por **ExGCs** pertenecen a una mayor diversidad de procesos celulares y cada una de las subunidades de proteínas multiméricas son codificadas, al menos parcialmente, por tripletes del tipo Ex1 o Ex2 (Palacios-Pérez and José, 2019).

En particular, las moléculas del proceso de traducción estaban ya presentes antes de LUCA, como las aminoacil-tRNA ligasas (aaRL) o la mayoría de las proteínas ribosomales, así como factores de traducción y proteínas de procesamiento de RNAs. También ya había algunas proteínas de la formación de pasajes de comunicación con el medio (transportadores) y proteínas de transducción de señales intracelulares (receptores y reguladores); así como algunas proteínas involucradas en el plegamiento o degradación de péptidos y proteínas; e involucradas en la síntesis o degradación de pared y membrana celulares. Por otro lado, varias de las proteínas de diversos metabolismos son también codificadas por ExGCs; de hecho, los procesos metabólicos más completos en esa etapa evolutiva estaban relacionados con el metabolismo energético y algunas proteínas constitutivas *-housekeeping-* del metabolismo de nt, aa, lípidos y carbohidratos. Tampoco es de sorprender que algunas proteínas del procesamiento del DNA son también codificadas por EGCs, ya sea que tuviesen un origen viral y/o que inicialmente sirvieran para la síntesis de RNA; finalmente, porciones de algunas proteínas de función aún indeterminada existían también antes del LUCA, lo que vuelve importante el investigarlas (Palacios-Pérez and José, 2019).

Del mismo modo que el código genético evoluciona desde su codificación primordial en RNY hacia dos vías distintas y complementarias, Ex1 y Ex2, el Ur-proteoma evolucionó desde los CSBSs a los **Proteomas Extendidos**. Observamos que la gran mayoría de las proteínas son codificadas por ambos tipos de tripletes, aunque las porciones codificadas por tripletes del tipo Ex2 son pocas y cubren una mayor extensión de las proteínas, en comparación con las cortas y numerosas porciones codificadas por tripletes del tipo Ex1, lo que podría deberse a que los tres codones que indican el frenado de la traducción de una proteína (codones de paro o *stop codons*) son codificados por Ex1. De esta manera, así como los codones de Ex1 y de Ex2 se complementan para dar lugar al SGC, las porciones codificadas por ambos ExGCs se complementan mutuamente, en la gran mayoría de los casos, para formar las proteínas que darán lugar al repertorio de LUCA, lo que puede observarse en la reconstrucción estructural particular de péptidos o proteínas codificadas por cada uno de los códigos genéticos (PGC o ExGCs). (Palacios-Pérez and José, 2019).

El primordio de las proteínas de los procesos más fundamentales puede rastrearse hasta su codificación por el PGC, además su codificación por Ex1 complementa su codificación por Ex2,

aunque en ambos casos se forma casi en su totalidad la proteína moderna; ejemplos de ello encontramos en las proteínas del proceso de traducción, como la proteína ribosomal S11 que desde el inicio unía cadenas ribonucleotídicas, o la ThrRL que unía compuestos con treonina desde su codificación por ambos tipos de ExGCs; o bien enzimas fundamentales del metabolismo, como la triosa-fosfato isomerasa (TPI o TIM) que con total certeza ya unía el ligando 3-fosfoglicerato aún con una codificación del tipo Ex1 o Ex2; también algunas subunidades de la acetil-coenzima A sintasa (AcCoAs) tienen porciones codificadas por tripletes del PGC y de ambos ExGCs como la subunidad alfa, mientras que la subunidad gamma sólo tiene porciones codificadas por ExGCs. Por otro lado, hay algunas proteínas cuya codificación por tripletes de algún tipo de ExGC es mejor que con el otro tipo de ExGC. Una de esas proteínas es la triptofanil-tRNA ligasa, que es codificada casi en su totalidad por tripletes del tipo Ex1, pero es casi inexistente en una codificación por tripletes del Ex2, los que precisamente no codifican para el aminoácido triptófano y es plausiblemente la explicación de tal diferencia; otra proteína con codificación diferencial es la girasa reversa (reGyr), que actualmente se relaciona con el procesamiento de DNA, aunque el péptido codificado por tripletes RNY unía cationes metálicos, ésta enzima parece estar codificada casi en su totalidad por tripletes del Ex2, pero con una pobre codificación por tripletes del tipo Ex1; todo lo que indicaría que el encontrar proteínas codificadas por ExGCs, no es por artefactos de la metodología (Palacios-Pérez and José, 2019).

De acuerdo con nuestros resultados, posiblemente ninguna de las rutas metabólicas estaba completa antes del LUCA pese a la variedad de proteínas; porque, aunque algunas proteínas estaban casi totalmente formadas y ya unían los ligandos que actualmente unen, no todas se encontraban en ese estado de formación. Sin embargo, existen indicios acerca del estilo de vida de los organismos que poseían los proteomas codificados por ExGCs. Anteriormente se describió que la vía metanogénica Wood-Ljungdhal correspondía al metabolismo del LUCA y aunque tal escenario es plausible, no todas las enzimas de la ruta se encontraron en la comparación genómica referida (Weiss et al., 2016); con nuestra metodología si es posible identificar la mayoría de las proteínas de dicho metabolismo reductivo, en particular encontramos que las subunidades de la AcCoAs, enzima esencial (*key enzyme*) de la vía Wood-Ljungdhal, son codificadas diferencialmente

por ambos EGCs, tal cual se mencionó en el párrafo previo. Otra de las proteínas que nosotros encontramos fue la reGyr, enzima propia de organismos termófilos para catalizar el embobinado de su material genético (McInerney, 2016), que es codificada casi totalmente por tripletes del código Ex2. Ambas enzimas son ejemplo del tipo de proteínas codificadas por ExGCs y, por ende, son posiblemente indicios del tipo de organismos que existían antes del LUCA (Palacios-Pérez and José, 2019).

Probablemente las moléculas hereditarias fueron inicialmente cortas e inestables y por ende la función más relevante de los primeros péptidos sintetizados habría sido la estabilización de moléculas ya existentes (Eigen, 1971; Woese, 1998), tales como nucleótidos y sus derivados, aminoácidos prebióticos o pequeños péptidos, cationes metálicos o incluso moléculas de agua (Szathmáry, 1993; Gesteland et al., 2006; Yarus, 2011); de manera que los módulos proteicos primordiales actuaron como dominios de unión en vez de poseer funciones enzimáticas que en realidad el RNA llevaba a cabo *e.g.* la autocatálisis (Gesteland et al., 2006).

Probablemente, algunos de los primeros péptidos en un inicio no fueron codificados sino que pudieron polimerizarse prebióticamente (Agmon, 2016; Tamura and Schimmel, 2003); en un ambiente donde los lípidos, también prebióticos, formaban vesículas que encerraban cadenas ribonucleotídicas y podían dividirse al alcanzarse cierta densidad de RNA (Lancet et al., 2018; Cornell et al., 2019); además ciertas moléculas con porciones nucleotídicas e iones metálicos servían para la transferencia de electrones y la catálisis de procesos proto-enzimáticos (Wächtershäuser, 1988a; Jadhav and Yarus, 2002; Fedor, 2002; Sharov, 2009; Sharov, 2016; Bray et al., 2018). En tal escenario, algunos de los aa sintetizados prebióticamente (Miller, 1953; Lazcano and Bada, 2003) fueron los primeros en ser codificados (Trifonov, 2004) por tripletes RNY (Eigen and Schuster, 1978) tales como A, S, D o V; la mayoría de esos aa, aún cuando participan en reacciones catalíticas, unen ligandos metálicos (como Asp) o son meros espectadores, por ejemplo para conferir estabilidad electrostática (como Ser). Por otro lado, los aa que participan directamente en reacciones enzimáticas son sintetizados por tripletes del tipo Ex1 y Ex2, como His o Lys (Ribeiro et al., 2020).

Los primeros módulos o plegamientos en consolidarse pudieron ser o del tipo ferredoxina y del tipo Rossman, que pudieron evolucionar para facilitar la transferencia de electrones y la catálisis de un proto-metabolismo (Raanan et al., 2020); o bien homólogos de aaRLs clase I, que a su vez son homólogas de proteínas que unen NAD específicamente y homólogos de aaRLs clase II, que son homólogas de chaperoninas (Rodin et al., 2009); de cualquier manera, todos esos plegamientos están formados principalmente por aa codificados por tripletes del tipo RNY y aparecen casi completamente codificados por tripletes de ambos ExGCs. Posteriormente, las primeras proteínas en consolidarse fueron probablemente las ribosomales, que inicialmente evolucionaron como chaperonas para dar estabilidad estructural a las cadenas ribonucleicas; la conformación inicial de dichas proteínas ribosomales pudo ser la base para la posterior codificación de proteínas metabólicas como la TIM (Lupas and Alva, 2017) y el primordio de todas ellas se remonta al PGC aunque están virtualmente formadas del todo por ambos ExGCs. De hecho, algunas de las proteínas codificadas por Ex1 y Ex2 ya se habían identificado entre las más antiguas (De Farias et al., 2016; Lupas and Alva, 2017) y se habían asignado al metabolismo de LUCA (Goldman et al., 2013; Weiss et al., 2016) e incluso al Primer Ancestro Común Universal o FUCA (por sus siglas en inglés), cuya existencia comprende desde el mundo del RNA hasta poco antes de la conformación final del LUCA (Prosdocimi et al., 2019; Prosdocimi and Farias, 2020).

Sabiendo las dos metodologías generales usadas para reconstruir rutas evolutivas, *bottom-up*, y *top-down* (Forterre and Gribaldo, 2007), recientemente se ha planteado que las propuestas más completas fusionan ambos tipos de enfoques (Mariscal et al., 2019; Preiner et al., 2020). Considero pues que ese es precisamente el tipo de planteamiento emprendido, ya que se reconstruyó la historia evolutiva del proteoma con base en la evolución del código genético, utilizando los genomas de muy diversas especies; planteamiento bioinformático basado en demostraciones puramente matemáticas (José et al., 2009; José et al., 2011). *Ergo*, iniciamos con las gráficas fenotípicas del **PGC** (José et al., 2015) y el **Ur-proteoma** codificado por éste (Palacios-Pérez et al., 2018), hasta la formación de los **Proteomas Extendidos** codificados por **ExGCs** y que por ende existieron antes del LUCA (Palacios-Pérez and José, 2019), cuyo proteoma ya estaba codificado por el SGC basado en DNA.

**6. EXTRA, OTROS PROYECTOS RELACIONADOS**

Adicional a la conformación de proteínas, otra manera de evaluar la evolución del fenotipo correspondiente al código genético, es observar las moléculas relacionadas, como los tRNAs de los tres dominios celulares (Zamudio et al., 2020); o el sitio de transferencia de peptidil (*peptidyl transferase centre*, PTC), en el que se lleva a cabo la traducción de proteínas (Prosdocimi et al., 2020); así como el RNA codificado por cada uno de los códigos genéticos descritos (Palacios-Pérez & José, en preparación).

Para analizar las secuencias de tRNAs y de PTCs, se ha utilizado el análisis de teoría de la información, que cuantifica la transmisión e intercambio de información a través de un canal o entre dos mensajes  $X$  y  $Y$ , de manera que se puede determinar la variación de la información compartida entre ellos *i.e.* la información mutua  $I(X,Y)$ . Esta medida  $I$  captura la información necesaria para describir una variable conociendo previamente la otra y todas aquellas con la misma variación de información se pueden agrupar entre sí (Zamudio et al., 2020).

LA TEORÍA DE LA INFORMACIÓN REVELA LA EVOLUCIÓN DE LOS ELEMENTOS DE IDENTIDAD DEL TRNA EN LOS TRES DOMINIOS DE LA VIDA

Por un lado, los tRNA poseen ciertos nucleótidos, elementos de identidad, que aseguran el cargado correcto de las aaRLs, indicando positiva y negativamente el aa correspondiente; dado que el sitio de activación de los tRNA es igual para todos los isoaceptores, el aa correcto está únicamente relacionado con el anticodón. En concreto se planteó si la teoría de la información puede ayudar a arrojar luz sobre la manera en que el anticodón quizá se relacione con otros sitios a lo largo de la secuencia de cada tRNA y, de esta manera, determinar elementos de identidad adicionales a los propuestos por otros grupos (Zamudio et al., 2020).

## LA HISTORIA ANTIGUA DE LA FORMACIÓN DEL PTC, CONTADA POR EL ANÁLISIS DE INFORMACIÓN Y DE LA CONSERVACIÓN DE SU SECUENCIA Y ESTRUCTURA

El PTC es el centro catalítico del ribosoma y forma parte del RNA ribosomal 23S (23SrRNA) y ha sido reconocido como la porción ribosómica más temprana, que se remonta al Primer Ancestro Común Universal (FUCA). Con frecuencia se asume que el PTC está altamente conservado en todos los seres vivos; sin embargo, no se ha dilucidado el grado de conservación de tal molécula a nivel secuencia y estructura, o las posibles diferencias intra- e inter- especies, además es posible que ciertos grupos de nt estén vinculados mutuamente a nivel informacional, hallazgos que podrían arrojar luz sobre la formación evolutiva del PTC mismo. Se descargaron entonces todas las secuencias completas disponibles del 23SrRNA de Bacteria y Arquea depositadas en GenBank, que se alinearon óptimamente por pares para recuperar la región PTC de 1424 secuencias de 23SrRNA, utilizando para ello la secuencia del PTC de *Thermus thermophilus* de 179 nt como cebador. Dichas secuencias de PTC se alinearon posteriormente entre sí y las regiones conservadas se asignaron, observaron y analizaron a lo largo de las estructuras primaria, secundaria y terciaria (Prosdocimi et al., 2020).

## LA EVOLUCIÓN DEL RNA PUEDE EXPLICAR LA EVOLUCIÓN DE LA TRADUCCIÓN

El origen y la evolución de la maquinaria de traducción es una de las principales transiciones evolutivas, aunque dicha maquinaria no puede entenderse sin la evolución del RNA, que bien pudo haberse configurado en un código genético primitivo con el patrón RNY y luego evolucionar mediante dos vías diferentes denominadas Extendidos 1 y 2, que finalmente dieron forma al código genético estándar actual. Con base en la ruta descrita, es posible rastrear la evolución del **RNAoma** (Palacios-Pérez & José, en preparación).

### **6.1 La teoría de la información revela la evolución de los elementos de identidad del tRNA en los tres dominios de la vida (colaboración)**

“We determined the identity elements of each tRNA isoacceptor for the three domains of life: Eubacteria, Archaea, and Eukarya. Our analyses encompass the most updated and curated available databases using an information theory approach. We obtained a collection of identity clusters for each of the isoacceptors of the 20 canonical amino acids for the three major domains of life. The identity clusters for all isoacceptors are compared within and among the three domains to determine their pattern of differentiation and to shed light on the evolution of the identity elements” (Zamudio et al., 2020).

### **6.2 La historia antigua de la formación del PTC, contada por el análisis de información y la conservación (colaboración)**

“The peptidyl transferase center (PTC) is the catalytic center of the ribosome and forms part of the 23S ribosomal RNA. The PTC has been recognized as the earliest ribosomal part and its origins embodied the First Universal Common Ancestor (FUCA). The PTC is frequently assumed to be highly conserved along all living beings. In this work, we posed the following questions: (i) How many 100% conserved bases can be found in the PTC? (ii) Is it possible to identify clusters of informationally linked nucleotides along its sequence? (iii) Can we propose how the PTC was formed? (iv) How does sequence conservation reflect on the secondary and tertiary structures of the PTC? Aiming to answer these questions, all available complete sequences of 23S ribosomal RNA from Bacteria and Archaea deposited on GenBank database were downloaded. Using a sequence bait of 179 bp from the PTC of *Thermus thermophilus*, we performed an optimum pairwise alignment to retrieve the PTC region from 1424 filtered 23S rRNA sequences. These PTC sequences were multiply aligned, and the conserved regions were assigned and observed along the primary, secondary, and tertiary structures. The PTC structure was observed to be more highly conserved close to the adenine located at the catalytical site. Clusters of interrelated, co-evolving nucleotides reinforce previous assumptions that the PTC was formed by the concatenation of proto-tRNAs and important residues responsible for its assembly were identified. The observed sequence variation does not seem to significantly affect the 3D structure of the PTC ribozyme” (Prosdocimi et al., 2020).

### **6.3 La evolución del RNA puede explicar la evolución de la traducción (autoría principal)**

“The origin and evolution of translational machinery is one of the major evolutionary transitions, although such machinery cannot be understood without the evolution of RNA. Such RNA could have been configured into a primeval genetic code, that follows an RNY pattern, that evolved by two different pathways dubbed Extended 1 and 2, that finally conformed the current standard genetic code.

We present here, how we trailed the evolutionary path of the **RNAome**, based on such described route. It was revealed that portions of the three types of ribosomal RNA (rRNAs) and portions of certain tRNA molecules were the first to being encoded and they configured as hairpins. Eventually, the complete rRNAs and tRNAs, as well as the RNA component of ribonuclease P, and the less-known molecules 6S RNA and the dual molecule tmRNA were encoded by triplets pertaining to both extended genetic codes. All that finally assembled cooperatively to each other, eventually resulting into the modern RNA molecules.” (Palacios-Pérez & José 2020, preparación).





# Information theory unveils the evolution of tRNA identity elements in the three domains of life

Gabriel S. Zamudio<sup>1</sup> · Miryam Palacios-Pérez<sup>1</sup> · Marco V. José<sup>1</sup>

Received: 21 April 2019 / Accepted: 3 September 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

## Abstract

We determined the identity elements of each tRNA isoacceptor for the three domains of life: Eubacteria, Archaea, and Eukarya. Our analyses encompass the most updated and curated available databases using an information theory approach. We obtained a collection of identity clusters for each of the isoacceptors of the 20 canonical amino acids for the three major domains of life. The identity clusters for all isoacceptors are compared within and among the three domains to determine their pattern of differentiation and to shed light on the evolution of the identity elements.

**Keywords** tRNA identity elements · Three domains of life · tRNA evolution · Information theory

## Introduction

The translation machine comprises an ample set of molecules interacting in a complex biological network. At the center of such network, it stands out the transfer RNA (tRNA) interacting with different molecules, including the messenger RNA (mRNA), the aminoacyl-tRNA synthetases (aaRSs), and the ribosome. Molecular recognition is a process that entails a principle of memory. To maintain such a complex interaction scheme, tRNAs possess two recognition codes in its structure, the anticodons for the mRNA, and the one known as “operational code” for the aaRS (De Duve 1988; Ribas de Pouplana and Schimmel 2001). The tRNAs operational code conducts the correct pairing of a tRNA with its cognate aaRS which has been previously charged with its corresponding amino acid. The attachment of the amino acid from the aaRSs to a tRNA is through an esterification reaction on the 3' end of the tRNA (Arnez and Moras 1997). After a tRNA has been correctly aminoacylated, it conducts the correct translation of the mRNA into a peptide through the ribosome. There are 20 aaRSs (one for each

amino acid of the standard genetic code), and each aaRS can be paired to a set of tRNAs with different anticodons; this set of tRNAs is known as isoacceptors. The existence of only 20 different aaRSs makes the operational code non-degenerate (Eriani et al. 1990). The family of the aaRSs enzymes are divided into two subfamilies (Class I and Class II) (Eriani et al. 1990). Both the operational code and the anticodon code did not evolve independently (Zamudio and José 2018; de Farias et al. 2018) since the early emergence of the mini-helix structure of the tRNA 3.5 billion years ago (Tamura 2015).

Modern aaRSs do not, in some cases, directly read the tRNA's anticodon (Ribas de Pouplana and Schimmel 2001). Although there is no explicit recognition of the anticodon, coupled with the degeneracy of the standard genetic code, tRNAs are charged with the correct amino acid. This correct aminoacylation of a tRNA is made through the operational code that is comprised by a set of identity elements that echo the information of the anticodon on the rest of the tRNA structure. Different methods have been used to identify the location of the identity elements, including experimental analysis (Giegé et al. 1998), and different mathematical and computational approaches (Zamudio and José 2018; Mukai et al. 2017; Branciamore et al. 2018). The consensus is that the set of identity elements differs for each isoacceptor group. The identity elements have been proposed to participate in the recognition of tRNAs not only in the aminoacylation reaction but also in the tRNA–protein interaction network (Ardell 2010). We have previously determined

✉ Gabriel S. Zamudio  
gazaso92@gmail.com

✉ Marco V. José  
marcojose@biomedicas.unam.mx

<sup>1</sup> Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, C.P. 04510 Mexico City, CDMX, Mexico

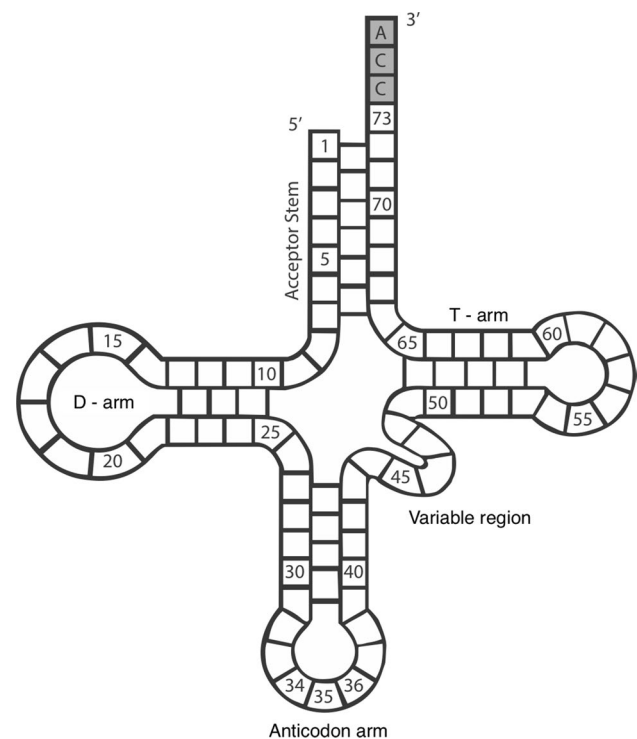
the identity elements regardless of the three main domains of life (Zamudio and José 2018). Differences have been found in the tRNA recognition system of the three domains of life (Woese et al. 2000). This collection of differences has been proposed to set a barrier of interdomain horizontal gene transfer of the aaRS genes (Ardell 2010). Differences in the aaRSs are also reflected in the tRNAs identities of each domain. tRNA genes have been found with different configurations in the three domains, such as multiple introns, split and tri-split tRNAs, and permuted tRNAs (Fujishima and Kanai 2014).

A tRNA has a canonical length of 72 nucleotides and is divided into four main sectors: the acceptor stem, the D-arm, the anticodon arm, and the T-arm and a sector known as variable region. Each arm is composed by a loop and a stem. In the 5' to 3' sense, the acceptor stem is joined to the D-arm followed by the anticodon arm; next is the variable region which connects to the T-arm that returns to the acceptor stem, thus forming a closed structure (Fig. 1). The 3' end of the tRNA is capped with an extra nucleotide and a terminal CCA which is added posttranscriptionally (Tamaki et al. 2018; Hou 2010). The D-arm has uridines modified nucleotides to dihydrouridines (Motorin and Grosjean 2005), and the T-arm is also known as *T $\Psi$ C*-arm due to the presence of thymidine, pseudouridine, and cytidine nucleotides. The variable region gets its name from the variable length it possesses.

We posed the question on how different the compendiums of the operational codes in the three main domains of life are. A hallmark of aaRSs is the exquisite specificity with which they select and aminoacylate only their cognate tRNA (Hendrickson 2001). Therefore, the discernment of the different operational codes becomes a central issue for a better understanding of the evolution of the translation system. In this work, the identity elements of the tRNAs are determined for each of the three domains of life using an information theoretical approach. We perform systematic analyses using the most updated and curated available databases (Jühling et al. 2009; Abe et al. 2014) (accessed August 2018). We derive a collection of identity clusters for each of the isoacceptors of the 20 canonical amino acids for the three major domains of life. The identity clusters for all the isoacceptors are compared within and among the three domains not only to determine their pattern of differentiation but also to gain insights on the origins and evolution of the identity elements and the translation system.

## Data sources

Data of mature nuclear tRNA gene sequences were downloaded from Jühling et al. (2009) and Abe et al. (2014) (accessed August 2018) for the 20 canonical amino acids.



**Fig. 1** Secondary structure of a tRNA with canonical length. The acceptor stem consists of the bases from 1 to 7 and from 66 to 72. The D-arm is constituted by bases 10 up to 25. The anticodon arm comprises the portion of bases from positions 27 to 43; the anticodon triplet is bases 34, 35, and 36. The variable region is made by bases from positions 44 to 48. The T-arm starts at base 49 and ends at position 65. The segment made by bases 8 and 9 connects the acceptor stem to the D-arm, while base 26 joins the D-arm with the anticodon arm. Base 73 and terminal CCA are added posttranscriptionally

The variable region of the sequences was removed to make the sequences comparable in length. The dataset was divided according to the three domains of life: Archaea, Eubacteria, and Eukarya, for the 20 tRNA isoacceptors. Two distinct databases were constructed from the tRNA sequences, due to the different lengths of the D-arm. In the first dataset, the D-loop was removed, while maintaining its respective stems (non-loop). For the second dataset, the length of the D-loop was determined per sequence and those whose length has more samples were considered (with loop). Duplicated sequences in each dataset were removed. The canonical length of the D-loop is eight nucleotides; however, in some cases, only one sequence exists, so the use of the canonical length was discarded in the analysis. The number of extra bases considered in the D-loop for the with-loop dataset is 1, 1, and 0, for Archaea, Eubacteria, and Eukarya, respectively. On the three domains, the isoacceptors sets for initiator methionine (ini-Met) and elongator methionine were differentiated according to the comments on the downloaded data. Isoacceptors sets with 15 or less sequences were not considered due to the low sample size.

## Methods

Information theory quantifies the transmission and sharing of information through an information channel or between two messages. Given two messages, it is possible to determine the information shared by them. The variation of information is pseudometric (the distance between any two different elements can be zero), and it measures the information distance between two messages  $X$  and  $Y$ . The variation of information is given by  $V(X, Y) = H(X) + H(Y) - 2I(X, Y)$ , where  $H(X)$  is the entropy of the random variable  $X$  and  $I(X, Y)$  is the mutual information between the two random variables  $X$  and  $Y$  (Zamudio and José 2018; Meilă 2003). For the analysis of tRNA sequences, we define as a continuous random variable the nucleotides in a given site on the set of tRNA isoacceptors for each amino acid. This allows us to compute the information distance between two sites in the tRNA structure within the isoacceptor groups on each domain. If any two given sites on the tRNA have an information distance of 0, then the occurrences of bases in the two sites are completely predictable one from the other. The variation of information allows clustering sites according to the variation of information among them. On any given cluster, all the sites on it will have a variation of information less or equal to a given parameter  $d$ . By setting the parameter to some specific values, such as  $d = 0$ , the clusters are well defined, whereas with a positive parameter the clusters can in some cases be fuzzy, i.e., a site on a tRNA isoacceptor may belong to two or more identity clusters. The appearance of fuzzy clusters is due to the triangle inequality property of any metric function such as the variation of information. For every isoacceptor in each domain, the value of maximum variation of information  $d_{\text{MAX}}$ , which ensures that for all the values  $d_1$ , such that  $0 \leq d_1 \leq d_{\text{MAX}}$ , the clusters inferred using the parameter  $d_1$  are well defined, was found. The clusters of sites formed with this parameter  $d_{\text{MAX}}$  of maximum information distance within each other comprise the collection of identity elements of each isoacceptor.

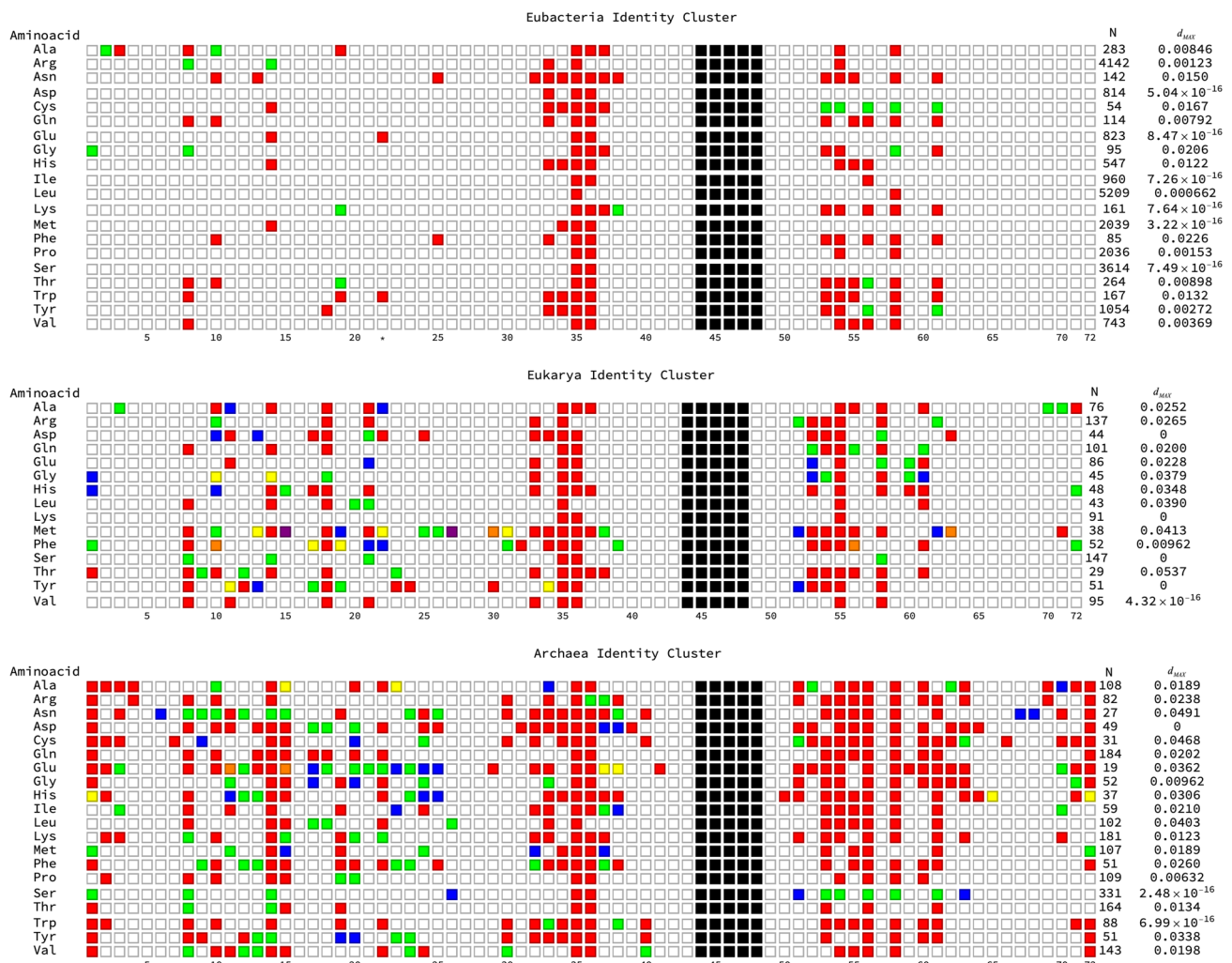
The use of the value  $d_{\text{MAX}}$  for the definition of the identity clusters determines the maximum value that allows the construction of well-defined clusters on each tRNA isoacceptor, albeit the information distance between any two sites of the same identity cluster  $d'$ , may be lower than  $d_{\text{MAX}}$ , i.e.,  $d' \leq d_{\text{MAX}}$ .

The supreme possible value of variation of information between any two clusters occurs when the variables  $X$  and  $Y$  are independent and uniformly distributed; in this case, the maximum value is  $d = 4$ , so in general the inequality  $0 \leq d \leq 4$  holds.

## Results

The variation of information for the three domains of life on the two datasets is computed. The positive value  $d_{\text{MAX}}$  for well-defined clusters was found, and several clusters of sites with an information distance lower than  $d_{\text{MAX}}$ , between the positions in each set are found for all the isoacceptors in the three domains for the with-loop dataset (Fig. 2) and for the non-loop dataset (“Appendix”). In Fig. 2, positions marked in red correspond to the set of sites related to the central anticodon base 35. Sets in different colors correspond to clusters of sites whose information distance is also lower than  $d_{\text{MAX}}$ . For the sake of clarity, pairs of sites corresponding to Watson–Crick pairs in the tRNA molecule that are at information distance lower than  $d_{\text{MAX}}$  are not marked. Positions marked in black correspond to the removed variable loop in both datasets, and the D-loop removed in the non-loop dataset. The numbering of the positions starts at the 5' end of the tRNA, and extra bases are marked with an asterisk (\*), so that the second half of the D-stem begins at position 22. The value of the parameter  $d_{\text{MAX}}$  in Fig. 2 is normalized to the maximum value of 4 by using the value  $d_{\text{MAX}}/4$ . After the normalization, the distance values lie on the interval  $0 \leq \frac{d_{\text{MAX}}}{4} \leq 1$ . In some isoacceptors, the information distance, which defines the identity clusters, is almost zero  $\approx 10^{-16}$ , which shows that the sites on the identity clusters are highly conserved on the isoacceptor with the exception of a tiny amount of rare mutations present in the dataset. For the rest of the isoacceptors, the higher values of information distance show that the sites are not highly conserved for a fixed nucleotide; instead, the nucleotides at the sites of an identity cluster follow a changing pattern which is predictable and thus conserved.

For archaeal tRNAs, the first four positions of the terminal side of the acceptor stem are present in clusters of sites related to the anticodon or with other positions in both datasets. Neither the bacterial nor the eukaryal tRNAs exhibit this pattern, albeit there are some exceptions. Positions 8 and 9 bridge the acceptor stem with the D-arm. Position 8 has a stronger presence in the identity clusters of all isoacceptors than position 9; in the cases where position 8 is associated with an identity cluster, it is usually associated with cluster containing the anticodon. For bacterial tRNAs, position 8 is present on some identity cluster for seven isoacceptors (Ala, Arg, Gln, Gly, Thr, Trp, and Val) in the with-loop dataset, from which Ala, Gly, Thr, and Val are amino acids found in Miller’s experiment (1953, 1957, 1974). For the Eukarya domain, position 8 is found in eight isoacceptors on both datasets. For the Archaea domain, position 8 is constantly absent from the identity clusters on the non-loop dataset with the



**Fig. 2** Identity clusters derived from the with-loop dataset. Extra sites in the D-loop are marked with asterisk (\*) after the end of the canonical D-loop at position 21. Sites removed of the variable region are in black. On each isoacceptor, the different colors (red, blue, green, yellow orange, purple) refer to the disjoint identity clusters grouping the sites of the each tRNA isoacceptor. The principal cluster containing the central anticodon base is colored in red. The value N stands

for the sample size of each isoacceptor group. The  $d_{MAX}$  value on the isoacceptors is the information distance that determines the identity clusters for each isoacceptor. The  $d_{MAX}$  value shown has been normalized to the maximum number of information distance of 4, as  $\frac{d_{MAX}}{4}$ . A similar figure derived for the non-loop dataset is provided in “Appendix” (color figure online)

exceptions of Gln and Pro (on which it is associated with the anticodon), and also for Asn and Glu. In contrast, position 8 is widely present on some identity clusters on the with-loop dataset. The D-stem is composed by positions 10–13 and 22–25. For the non-loop dataset of Eubacteria and Eukarya, there are some cases with clusters conformed only by bases forming Watson–Crick pairings, and only position 10 of Phe in Eukarya is part of a wider cluster. An ample number of identity elements appear in the D-stem for the archaeal non-loop dataset. In the with-loop dataset, bacterial tRNAs display a number of identity elements along the D-arm which increases for Eukarya and Archaea. For the archaeal D-loop, there is a constant occurrence of position 14 as being part of an identity cluster. This

occurrence is less present in Eukarya, while for Eubacteria position 14 is contained in an identity cluster in only five isoacceptors. The anticodon arm is composed of bases 27–43, with position 26 connecting the D-arm with the anticodon arm. In the three domains, for both datasets, the first three base pairs of the anticodon stem, to wit, 27–43, 28–42, and 29–41, in general do not belong to any wide identity cluster, with two exceptions in the with-loop dataset. Such exceptions are the isoacceptor for Glu in the Archaea whose domain has the pair 29–41 related to the anticodon and for Met in Eukarya where position 27 is related to position 15. The anticodon loop in all domains possesses bases that are part of the identity cluster associated with the central position 35 of the anticodon. Note

that in some cases, the loop bases surrounding the anticodon triplet are part of an identity cluster which is not associated with the anticodon. The T-arm is composed of the positions 49–65 in the canonical structure. The tRNAs for Archaea present a wider spectrum of identity elements in T-arm for both datasets than the other two domains. The clusters of identity elements for the non-loop Eukarya tRNAs have a wider presence in the T-stem than in the corresponding bacterial tRNAs; meanwhile, on the with-loop dataset, the domains of Eukarya and Eubacteria have an average of equal identity elements across the T-arm. Position 57, which is the central position of the T-loop, does not belong to any identity cluster in any domain, while the surrounding bases are contained in a cluster in most cases.

## Discussion

In this work, we calculated the identity elements of the 20 tRNA canonic isoacceptors for Eubacteria, Archaea, and Eukarya domains using a metric from information theory. Current datasets contain tRNA gene sequences with different lengths of the D-loop; hence, sequences were analyzed with and without the D-loop. An updated catalog of the identity elements for each of the main domains of life is presented. The purpose of using two datasets was to determine whether the length of the D-loop is associated with the identity elements of the operational code in tRNAs. Subtle differences in the identity elements appeared in the D-stems that show that, in some cases, the D-loop influences the operational code. The variable region was deleted in order to make the tRNA sequences generally equal in length. Different identity clusters are found for each isoacceptor on the three domains. In the three domains, the with-loop isoacceptors present an increased number of identity elements in the D-stem than their corresponding counterparts of the non-loop dataset. A general pattern appears that resembles the phylogenetic tree of the three domains. Most of the isoacceptors in the three domains present identity elements in the T-arm, while tRNAs from Eukarya and Archaea possess a higher number of identity elements in the D-arm when compared to Eubacteria. Finally, archaeal isoacceptors show identity elements in the acceptor stem, which are absent in the other two domains. There is a high bias in repositories of bacterial tRNA sequences; Eukarya and Archaea have sample sizes of the same magnitude. The use of a positive value  $d$  as the parameter of the variation of information allows for the analysis of not completely strict similarity patterns. The variation of information between two sites, as information metric, gives the information necessary to discern the value of one parameter given the knowledge of the other. A variation of information of 0 between two sites results in the need

of no more information to determine the value of one site from previously knowing the other. The clustering parameter,  $d_{MAX}$ , determines the magnitude of the variability for all the identity clusters in any given isoacceptor.

The widely established base pair G3:U70 determinant in tRNA<sup>Ala</sup> (Ribas de Pouplana and Schimmel 2001; Hou and Schimmel 1988; McClain and Foss 1988; Chong et al. 2018) is part of an identity cluster in the three domains. In bacterial tRNAs<sup>Ala</sup>, base 3 is on the same cluster as the anticodon; for the Eukarya domain, a cluster conformed by bases 3, 70, and 71 is formed; in Archaea, base 3 is related to the anticodon central position, while base 70 is related to base 33 which delimits the anticodon triplet. It has been reported that eukaryal AlaRS has gained functionality by mischarging non-cognate tRNAs due to the recognition of the pair G4:U69 (Sun et al. 2016). Archaeal AlaRS possesses the same mechanisms for detecting the G4:U69 base pair in non-cognate tRNAs (Sun et al. 2016); this base pair is present in the anticodon identity cluster for tRNAs<sup>Ala</sup> in the Archaea domain. Mechanisms for correction of the mischarging of tRNA<sup>Thr</sup> with alanine by AlaRS in kingdom Animalia have been described (Kuncha et al. 2018). The G:U wobble base pair is a fundamental unit of RNA secondary structure in all three phylogenetic domains (Varani and McClain 2000).

Differences with the identity clusters found previously arise (Zamudio and José 2018). This is a consequence of using the three domains in the same dataset and restricted the analyses to sequences with the D-loop of eight nucleotides, which is the canonic length, coupled with the use of an information distance that allows variability.

A distinctive feature of tRNA<sup>His</sup> is an extra 5' nucleotide that is usually a guanylate at position G:-1 (Wang et al. 2007). This nucleotide is added posttranscriptionally, and therefore, it was not included in our gene analysis.

Some bases in the identity clusters correspond to positions associated with posttranscriptional modifications which have been reported to be either universal in the three domains of life (Jühling et al. 2009), generally present in two domains, or are domain specific (Motorin and Grosjean 2005; Lorenz et al. 2017). Modifications of nucleobases from posttranscriptional modifications enhance the stability of tRNAs and improve its interaction with other molecules involved in translation, such as aaRS, translation factors, or the mRNA (Motorin and Grosjean 2005).

The D-arm receives its name as it contains the modified base dihydrouridine (Lorenz et al. 2017), which is the result of adding two hydrogen atoms to a uridine nucleoside. The D-arm provides structural stability to the tRNA and avoids its premature dissociation from the ribosome (Smith and Yarus 1989). Such a degree of interaction between the D-arm with aaRS is more notorious in bacterial and eukaryal tRNAs, whereas such arm does not seem imperative for

Archaea (Tamaki et al. 2018), which could help to explain the identity clusters and its dendrograms. However, the D-arm confers a more precise differentiation between each other tRNA. On the D-loop, the positions associated with the anticodon are not the ones which are modified to dihydrouridines, which provide flexibility to this region (Motorin and Grosjean 2005). One of these bases is position 14 which belongs to an identity cluster in some bacterial tRNAs, half of the eukaryal isoacceptors, and it is a general property of archaeal tRNAs and is not modified to dihydrouridine. Positions of tRNA anti-determinant bases, i.e., positions on which the presence of a specific base disassociates the recognition of the tRNA with its corresponding aaRS, are not generally discernible by our methodology. Such is the case of C:34 for bacterial tRNA<sup>Ile</sup> which is not present on any identity cluster; the opposite example is position 10 for eukaryal tRNA<sup>Phe</sup>. This position has been reported as an important base for the recognition with its corresponding aaRS on yeast (Motorin and Grosjean 2005). In addition, this position arises as an identity element associated with position 56 on the T-loop. Positions 10 and 56 are on opposite sides of the tRNA on the cloverleaf representation and are arranged on opposite sides at the corners in the L-shaped tertiary structure. The nucleotides on positions 8 and 10 undergo modifications to thiouridine for photon protection (Motorin and Grosjean 2005) and N2-methylguanosine for proper folding (Lorenz et al. 2017), respectively. Position 8 is related to the anticodon on some bacterial tRNAs, whereas position 10 is an identity element for some eukaryal and archaeal tRNAs. Our results suggest intricate relationships between the positions related to the structural properties and preservation and the identity elements of the tRNAs. Giegé stated that identity elements are rare in D-arm; however, some determinants in the tRNAs for certain amino acids have been characterized (Hendrickson 2001). Herein, we report that identity elements in the D-loop of Archaea and Eukarya are ubiquitous. For example, position 20 in bacteria does not appear in any identity cluster, whereas in Archaea this position is fixed in the 20 amino acids and in Eukarya appears in eight amino acids. It has been suggested the existence of multiple sets of identity elements for each isoacceptor (Giegé et al. 1998), and this has been corroborated (Zamudio and José 2018).

The positions in the anticodon loop adjacent to the anticodon triplet are subject to modifications of the nucleosides in order to ensure accuracy and efficiency in translation (Motorin and Grosjean 2005); this property is reflected in the fact that in several isoacceptors the positions around the anticodon are in the same cluster as the anticodon central base. Likewise, according to reported tRNA molecule sequences (Jühling et al. 2009), position 57, in the

canonical structure length, remains unmodified in the vast majority of organisms and this base is consistently unassociated with any identity cluster in the three domains of life. Positions 55 and 58 are modified to pseudouridine and 1-methyladenosine, respectively, in the three domains of life, and both positions are associated with an identity cluster, generally the cluster for the anticodon, on most tRNAs of the three domains. For archaeal tRNAs, positions 55 and 58 are associated with base 56, and these three positions could indicate major recognition sites between the aaRSs and its cognate tRNA. On the D-loop, the positions associated with the anticodon are not the ones which are modified to dihydrouridines, which provide flexibility to this region (Motorin and Grosjean 2005). One of these bases is position 14 which belongs to an identity cluster in some bacterial tRNAs, half of the eukaryal isoacceptors, and it is a general property of archaeal tRNAs and is not modified to dihydrouridine. Positions of tRNA anti-determinant bases, i.e., positions on which the presence of a specific base disassociates the recognition of the tRNA with its corresponding aaRSs, are not generally discernible by our methodology. Such is the case of C:34 for bacterial tRNA<sup>Ile</sup> which is no present on any identity cluster; the opposite example is position 10 for eukaryal tRNA<sup>Phe</sup>; this position has been reported as an important base for the recognition with its corresponding aaRSs on yeast (Motorin and Grosjean 2005). This position arises as an identity element associated with position 56 on the T-loop. Positions 10 and 56 are on opposite sides of the tRNA on the cloverleaf representation and are arranged on opposite sides of the corner in the L-shaped tertiary structure. The nucleotides on positions 8 and 10 undergo modifications to thiouridine for photon protection (Motorin and Grosjean 2005) and N2-methylguanosine for proper folding (Lorenz et al. 2017), respectively. Position 8 is related to the anticodon on some bacterial tRNAs, whereas position 10 is an identity element for some eukaryal and archaeal tRNAs. Our results suggest intricate relationships between the positions related to the structural properties and preservation and the identity elements of the tRNAs.

We remark that when comparing the information theory approach with a conservation analysis, the information theory extends the results from a conservation analysis. If a base is fully conserved in a tRNA sequence, it will also be shown by the information theory approach because its variation of information is zero. If a base is less conserved, then its variation of information will be greater than zero. The less conserved a base is, relative to the others, the more will increase its variation of information.

The anticodon and acceptor arms are fully conserved but that is not the case with the variable arm. Whereas in middle range of conservation it would be the D and T

arms, D-arm confers the highest interaction with an aaRSs of the two (Tamaki et al. 2018). In agreement with the structural analysis by Tamaki et al. (2018), we found that positions 8, 10, 14, 19, 33, 53, 54, 55, 56, 58, and 61 are mostly conserved in the three domains of life.

Identity elements have functions beyond the correct interaction of a tRNA with its corresponding aaRS. The operational code also plays a role in guiding the correct folding of the tRNA to its tertiary structure. This property is manifested with the cluster of bases 10 and 56 that is present in Phe of Eukarya. Bases 10 and 56 belong to the D-loop and the T-loop, respectively, and they come into contact when the tRNA folds into its tertiary structure. In contrast to Giegé et al. (1998), we found tRNA identity nucleotides in the anticodon loop of bacterial tRNA<sup>Leu</sup>, tRNA<sup>Ser</sup>, and tRNA<sup>Ala</sup>.

The D-arm receives its name as it contains the modified base dihydrouridine (Lorenz et al. 2017), which is the result of adding two hydrogen atoms to a uridine nucleoside. The D-arm provides structural stability to the tRNA and avoids its premature dissociation from the ribosome (Smith and Yarus 1989). Such a degree of interaction between the D-arm with aaRS is more notorious in bacterial and eukaryal tRNAs, whereas such arm does not seem imperative for Archaea (Tamaki et al. 2018), which could help to explain the identity clusters and its dendrograms. However, the D-arm confers a more precise differentiation between each other tRNA. Giegé stated that identity elements are rare in D-arm; however, some determinants in the tRNAs for certain amino acids have been characterized (Hendrickson 2001). Herein, we report that identity elements in the D-loop of Archaea and Eukarya are ubiquitous. For example, position 20 in bacteria does not appear in any identity cluster, whereas in Archaea this position is fixed in the 20 amino acids and in Eukarya appears in eight amino acids. It has been suggested the existence of multiple sets of identity elements for each isoacceptor (Giegé et al. 1998), and this has been corroborated (Zamudio and José 2018). Each tRNA molecule is usually composed by the D-arm, the anticodon arm, the variable region, the T-arm, and the acceptor stem with the addition of the terminal CCA added posttranscriptionally (Hou 2010). There are some few D-armless organisms and even one species described with only the anticodon and acceptor regions (Fujishima and Kanai 2014). The high number of identity elements in Archaea may provide robustness to the recognition system, given that several types of disruption of tRNA molecules have been observed, while still maintaining its functionality (Fujishima and Kanai 2014).

Carter and Wills (2018) have performed regression analyses of different qualitative features of bacterial

tRNAs, revealing that the acceptor stem of bacterial tRNAs retains an ancient operational code based on thermodynamic attributes, which are recognized by the corresponding aaRS. Our work provides a way to detect identity elements based on the mutual information of the anticodon with respect to other sites throughout the tRNA molecule, i.e., we look how the information contained in the anticodon could be reflected in other positions, which can be recognized by the aaRS not “seeing” the anticodon; our work is independent of structural considerations, which seems to be of high importance for bacterial tRNAs (Carter and Wills 2018), for which our approach does not detect many identity elements, in contrast to more complex or extremophile organisms such as Archaea, that usually have minimal tRNAs. Therefore, the abundance of identity elements may be necessary to guarantee the correct aminoacylation.

A wide number of functions of uncharged tRNAs outside the translation framework have been recently found. These roles include gene regulation, degradation, and cellular apoptosis (Raina and Ibba 2014). The identity elements in each isoacceptor could be related to the interaction of tRNAs in different biological networks.

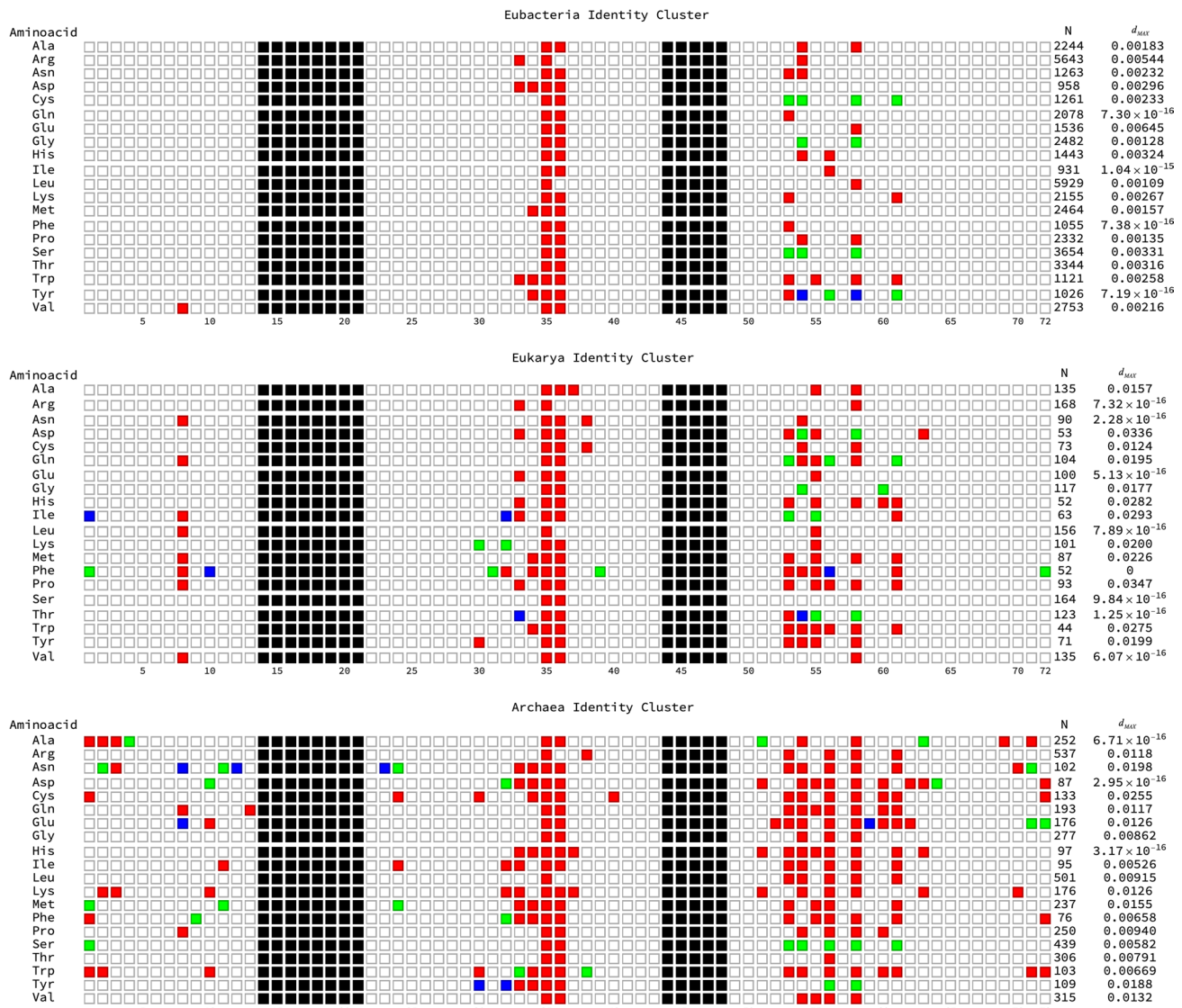
We were able to use the information theory to capture the evolution of the operational code; particularly, we tracked the information associated with the anticodon throughout the whole tRNA structure, which could help to explain more accurately how the aaRS can charge the correct amino acid without “seeing” the anticodon. This kind of approach would certainly be robust with the repository of more and diverse archaeal tRNAs. Additionally, the combination of different methods could improve the elucidation of the operational code and its evolution.

**Acknowledgements** We thank Juan R. Bobadilla for technical computer support. We thank the anonymous reviewer for their helpful criticisms and suggestions. We thank Adhemar Liquitaya-Montiel for helpful discussions at the beginning of this work.

**Authors' contributions** GSZ and MVJ conceived and designed the experiments; GSZ performed the experiments and analyzed the identity clusters; MPP analyzed tRNA sites for posttranscriptional modifications; GSZ, MPP, and MVJ analyzed the data; GSZ and MVJ wrote the paper.

**Funding** GSZ is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM), and received a doctoral fellowship from CONACYT (Number: 737920). MP is a doctoral student from Programa de Doctorado en Ciencias Biomédicas (PDCB), Universidad Nacional Autónoma de México (UNAM), and she receives the fellowship 694877 from CONACYT. MVJ was funded by Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México, UNAM (PAPIIT-IN201019).

## Appendix



## References

- Abe T, Inokuchi H, Yamada Y, Muto A, Iwasaki Y, Ikemura T (2014) TRNADB-CE: tRNA gene database well-timed in the era of big sequence data. *Front Genet* 5:114
- Ardell DH (2010) Computational analysis of tRNA identity. *FEBS Lett* 584:325–333
- Arnez JG, Moras D (1997) Structural and functional considerations of the aminoacylation reaction. *Trends Biochem Sci* 22:211–216
- Branciamore S, Gogoshin G, Di Giulio M, Rodin A (2018) Intrinsic properties of tRNA molecules as deciphered via bayesian network and distribution divergence analysis. *Life* 8:5
- Carter CW, Wills PR (2018) Hierarchical groove discrimination by Class I and II aminoacyl-tRNA synthetases reveals a palimpsest of the operational RNA code in the tRNA acceptor-stem bases. *Nucleic Acids Res* 46:9667–9683
- Chong YE, Guo M, Yang X-L, Kuhle B, Naganuma M, Sekine S-I, Yokoyama S, Schimmel P (2018) Distinct ways of G: U recognition by conserved tRNA binding motifs. *Proc Natl Acad Sci* 115:7527–7532
- De Duve C (1988) The second genetic code. *Nature* 333:117–118
- de Farias ST, Antonino D, Rêgo TG, José MV (2018) Structural evolution of Glycyl-tRNA synthetases alpha subunit and its implication in the initial organization of the decoding system. *Prog Biophys Mol Biol* 30:1e8
- Eriani G, Delarue M, Poch O, Gangloff J, Moras D (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* 347:203–206
- Fujishima K, Kanai A (2014) tRNA gene diversity in the three domains of life. *Front Genet* 5:142






- Giegé R, Sissler M, Florentz C (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res* 26:5017–5035
- Hendrickson TL (2001) Recognizing the D-loop of transfer RNAs. *Proc Natl Acad Sci* 98:13473–13475
- Hou Y-M (2010) CCA addition to tRNA: implications for tRNA quality control. *IUBMB Life* 62:251–260
- Hou Y-M, Schimmel P (1988) A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* 333:140–145
- Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* 37:D159–D162
- Kuncha SK, Mazeed M, Singh R, Kattula B, Routh SB, Sankaranarayanan R (2018) A chiral selectivity relaxed paralog of DTD for proofreading tRNA mischarging in Animalia. *Nat Commun* 9:511
- Lorenz C, Lünse C, Mörl M (2017) tRNA modifications: impact on structure and thermal adaptation. *Biomolecules* 7:35
- McClain WH, Foss K (1988) Changing the identity of a tRNA by introducing a G-U wobble pair near the 3' acceptor end. *Science* (80-) 240:793–796
- Meilä M (2003) Comparing clusterings by the variation of information. In: *Proceedings of learning theory and kernel machines: 16th annual conference on learning theory and 7th kernel workshop, COLT/Kernel 2003, Washington, DC, USA, August 24–27, 2003*. Springer, Berlin, pp 173–187
- Miller SL (1953) A production of amino acids under possible primitive earth conditions. *Science* (80-) 117:528–529
- Miller SL (1957) The mechanism of synthesis of amino acids by electric discharges. *Biochim Biophys Acta* 23:480–489
- Miller SL, Orgel LE (1974) *The origins of life on the earth*. Prentice-Hall, Upper Saddle River
- Motorin Y, Grosjean H (2005) tRNA modification. In: *Encyclopedia of life sciences*. Wiley. <https://doi.org/10.1038/npgs.els0003866>
- Mukai T, Reynolds N, Crnković A, Söll D (2017) Bioinformatic analysis reveals archaean tRNA<sup>Tyr</sup> and tRNA<sup>Trp</sup> identities in bacteria. *Life* 7:8
- Raina M, Ibba M (2014) tRNAs as regulators of biological processes. *Front Genet* 5:171
- Ribas de Pouplana L, Schimmel P (2001) Operational RNA code for amino acids in relation to genetic code in evolution. *J Biol Chem* 276:6881–6884
- Smith D, Yarus M (1989) Transfer RNA structure and coding specificity. I. Evidence that a D-arm mutation reduces tRNA dissociation from the ribosome. *J Mol Biol* 206:489–501
- Sun L, Gomes AC, He W, Zhou H, Wang X, Pan DW, Schimmel P, Pan T, Yang XL (2016) Evolutionary gain of alanine mischarging to noncognate tRNAs with a G4:U69 base pair. *J Am Chem Soc* 138:12948–12955
- Tamaki S, Tomita M, Suzuki H, Kanai A (2018) Systematic analysis of the binding surfaces between tRNAs and their respective aminoacyl tRNA synthetase based on structural and evolutionary data. *Front Genet* 8:227
- Tamura K (2015) Origins and early evolution of the tRNA molecule. *Life* 5:1687–1699
- Varani G, McClain WH (2000) The G-U wobble base pair. *EMBO Rep* 1:18–23
- Wang C, Sobral B, Williams K (2007) Loss of a universal tRNA feature. *J Bacteriol* 189:1954–1962
- Woese CR, Olsen GJ, Ibba M, Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* 64:202–236
- Zamudio GS, José MV (2018) Identity elements of tRNA as derived from information analysis. *Orig Life Evol Biosph* 48:73–81

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Article

# The Ancient History of Peptidyl Transferase Center Formation as Told by Conservation and Information Analyses

Francisco Prosdocimi <sup>1,2,\*</sup> , Gabriel S. Zamudio <sup>2</sup> , Miryam Palacios-Pérez <sup>2</sup>,  
Sávio Torres de Farias <sup>3</sup> and Marco V. José <sup>2,\*</sup> 

<sup>1</sup> Laboratório de Biologia Teórica e de Sistemas, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21.941-902, Brazil

<sup>2</sup> Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Ciudad Universitaria, CDMX 04510, Mexico; gazaso92@gmail.com (G.S.Z.); mir.pape@iibiomedicas.unam.mx (M.P.-P.)

<sup>3</sup> Laboratório de Genética Evolutiva Paulo Leminsk, Departamento de Biologia Molecular, Universidade Federal da Paraíba, João Pessoa, Paraíba 58051-900, Brazil; stfarias@yahoo.com.br

\* Correspondence: prosdocimi@bioqmed.ufrj.br (F.P.); marcojose@biomedicas.unam.mx (M.V.J.)

Received: 7 July 2020; Accepted: 31 July 2020; Published: 5 August 2020



**Abstract:** The peptidyl transferase center (PTC) is the catalytic center of the ribosome and forms part of the 23S ribosomal RNA. The PTC has been recognized as the earliest ribosomal part and its origins embodied the First Universal Common Ancestor (FUCA). The PTC is frequently assumed to be highly conserved along all living beings. In this work, we posed the following questions: (i) How many 100% conserved bases can be found in the PTC? (ii) Is it possible to identify clusters of informationally linked nucleotides along its sequence? (iii) Can we propose how the PTC was formed? (iv) How does sequence conservation reflect on the secondary and tertiary structures of the PTC? Aiming to answer these questions, all available complete sequences of 23S ribosomal RNA from Bacteria and Archaea deposited on GenBank database were downloaded. Using a sequence bait of 179 bp from the PTC of *Thermus thermophilus*, we performed an optimum pairwise alignment to retrieve the PTC region from 1424 filtered 23S rRNA sequences. These PTC sequences were multiply aligned, and the conserved regions were assigned and observed along the primary, secondary, and tertiary structures. The PTC structure was observed to be more highly conserved close to the adenine located at the catalytical site. Clusters of interrelated, co-evolving nucleotides reinforce previous assumptions that the PTC was formed by the concatenation of proto-tRNAs and important residues responsible for its assembly were identified. The observed sequence variation does not seem to significantly affect the 3D structure of the PTC ribozyme.

**Keywords:** peptidyl transferase center; origin of life; 23S rRNA; proto-tRNA; emergence of biological systems

## 1. Introduction

The peptidyl transferase center (PTC) is the catalytic center of the ribosome. Being a specific region of the larger ribosomal subunit, it is responsible for binding activated amino acids together and performing peptide elongation during protein synthesis. Since the early 1980s, Carl Woese and Harry Noller noticed that the essential mechanism underlying translation might be RNA based [1]. Nevertheless, it was only in 1992 that Noller and his collaborators found experimental evidence to support the idea that the PTC was indeed a ribozyme. They confirmed that the activity of peptidyl transferase is held by the ribosomal RNA after treating ribosomes with proteases without prejudice

to the peptidic bond formation [2]. Four years later, Peter Lohse and Jack Szostak carried out the in vitro selection of ribozymes with the capability “to synthesize ester and amide linkages, as does the ribosomal peptidyl transferase” [3]. Further studies confirmed that the ribosomal peptidyl transferase reaction was performed by a region smaller than 200 base pairs located in the 23S ribosomal RNA of prokaryotes. Eukaryotes contain a similar PTC located in their 28S ribosomal RNA.

The PTC region has been considered crucial in the understanding about the origins of life. It has been described as the most significant trigger that engendered a mutualistic behavior between nucleic acids and peptides, allowing the emergence of biological systems [4–6]. Additionally, the proposal of a First Universal Common Ancestor (FUCA) departed from the contingent appearance of an ancient ribozyme capable of binding amino acids together [7]. The emergence of this proto-PTC is a prerequisite to couple a chemical symbiosis between RNAs and peptides that further evolved both to (i) become the large subunit of the ribosome by the principle of accretion and (ii) to allow the emergence of the genetic code. Although there remains controversy in the literature about whether the PTC is ancient or not [8–11], its importance cannot be challenged as it composes the central core of the decoding language of biology. The PTC is a versatile catalyst [12] that works as a turnstile for binding 20 different and very specific L-amino acids together to compose every cellular protein [13,14].

The origin and initial evolution of the PTC is a fertile field of debate and discussion in the scientific community. Some works indicated that the PTC was formed by a duplication of ancient forms of RNA once its structure was symmetric [15,16]. Other studies proposed the formation of the PTC by the junction of smaller RNAs, such as primitive tRNAs. Tamura [17] analyzed the secondary structure of the PTC and observed topological similarities with the secondary structure of transfer RNAs. In this sense, Caetano-Anollés and Sun [18] used structural analyses to provide evidence that tRNAs were older than ribosomes and were coopted to operate in the translation machinery. Farias and collaborators [19] analyzed sequence similarities between reconstructed ancestral sequences of tRNAs and the PTC. They verified an identity of 50.5% between a modern PTC and concatemers of ancestral tRNAs. Additionally, Root-Bernstein and Root-Bernstein [20] studied the similarity between tRNAs and rRNAs from *Escherichia*, observing several tRNA sequences found along its 23S rRNA sequence. They also suggested that the ribosomal RNA might have functioned as a primitive genome. Farias et al. [21] reconstructed a 3D structure of the PTC based on an ancestral sequence of tRNAs and observed a structural similarity of 92% when compared to the PTC of the bacteria *Thermus thermophilus*. Additionally, Demongeot and Seligmann [22] performed comparative studies between the secondary structure of both tRNAs and rRNAs and suggested that rRNAs were probably originated from tRNA molecules. Together, all these data make evident a scenario for the origin of life in which an evolutionary and chronological connection can be observed between these two essential components of the translation system: tRNAs and rRNAs.

Due to its remarkable relevance to biology and to the origins of life field, new studies that approach issues relating tRNAs and rRNAs are indispensable to better clarify how the initial organization of biological systems took place. Even when most works about ribosomal structure indicate that the PTC is highly conserved among all forms of life, we were unable to find conservation analyses of this particularly interesting region of the ribosome among the ancient domains of life. In addition, it seems important to analyze both the sequence and structure of the PTC in detail to gain insights about the emergence of biological systems. In this work, the following questions were posited: (i) Which exact nucleotides from the PTC are conserved among prokaryotes? (ii) How was the PTC probably formed? (iii) How can molecular modeling answer questions about the 3D structure conservation of the catalytic site of the ribosome? (iv) Are there co-evolving clusters of nucleotides that were invariant throughout the PTC's evolution? Herein, we used comparative genomics and information theory to unravel patterns of information variation and nucleotide conservation among PTCs using all complete sequences of 23S rRNAs available in the GenBank database [23].

## 2. Material and Methods

### 2.1. Download of Complete 23S Ribosomal RNAs from Public Databases

All available sequences (complete) of 23S rRNA were retrieved from GenBank using the following search: “23s ribosomal RNA [All Fields] AND complete [All Fields] AND biomol\_rrna [PROP]” with the nucleotide search function of the National Center for Biotechnology Information (NCBI) website. This search resulted in 1434 sequences downloaded from GenBank [23].

### 2.2. Retrieving PTCs from 23S Ribosomal RNA Sequences

A PTC sequence containing 179 bp from the bacteria *T. thermophilus* was obtained [19] and used as bait to retrieve the PTC region from the other 23S rRNA sequences obtained. The selection of PTC regions was performed using the optimal pairwise alignment tool Needleman–Wunsch [24]. Each of the 1434 23S rRNA sequences downloaded were aligned to the PTC of *T. thermophilus* using the needle script provided in the EMBOSS package [25]. An in-house needleParser.pl Perl script was developed to retrieve the start and end coordinates of the alignment and another script named get\_RegionByCoordinate.pl was used to retrieve the exact PTC from the 23S rRNA sequences obtained.

### 2.3. Multiple Alignment, Filtering, and Production of PTC Datasets

The 1434 PTC regions were multiple aligned using ClustalW [26] software. After visual inspection of the alignment, we noticed 10 sequences particularly divergent from the others, presenting exceeded nucleotides and possibly representing annotation errors. These sequences were filtered to provide a PTC dataset in FASTA format containing 1424 high-quality PTC regions.

The whole PTC dataset was also separated into three different subsampled datasets according to taxonomic information. We analyzed the whole dataset and sequences from two ancient domains of life: Bacteria and Archaea. Five PTC sequences from eukaryotic organisms were discarded from further analyses as it has been found that eukaryotes compose a derived clade originated from the Lokiarchaeota group of archaea from the subphylum Asgard [27,28]. Each domain dataset was analyzed separately. The conservation analysis of nucleotides was performed using sequence alignments and the WebLogo tool [29] was used to generate pictures identifying the most conserved nucleotide residues. Manual curation was also performed in the alignments in order to allow the identification of 100% conserved nucleotides.

### 2.4. Information Theory Analysis of PTCs

For the information theory analysis, we considered the pseudometric variation of information. Pseudometrics differ from metrics because two different elements can be at distance zero. The variation of information measures the distance between two messages,  $X$  and  $Y$ . This metric is given by the formula  $V(X,Y) = H(X) + H(Y) - 2I(X,Y)$ , where  $H(X)$  stands for the entropy of the random variable  $X$  and  $I(X,Y)$  is the mutual information shared between the two random variables  $X$  and  $Y$  [30,31]; both measures are considered in bit units. For the analysis of PTC sequences, we deduced from the alignment the discrete distribution of nucleotides at each position. This procedure allowed us to determine the information distance of any two sites on any set of sequences of the same length [30,32]. If two positions are at an information distance of 0, the occurrences of nucleotides at these positions are strictly predictable, i.e., it is possible to determine one nucleotide from the other. Note that this fact holds in conserved positions but also in the case when the sites present some sort of linked variation. The variation of information allowed us to cluster sites according to the information distance between them. Particularly, an intra-cluster information distance of 0 provided well-defined, non-fuzzy clusters, on which nucleotides within single clusters perfectly co-varied.

### 2.5. Mapping Conservation into 2D and 3D PTC Structures

The 2D structure of the PTC was obtained from Ribovision [33]. We downloaded the SVG picture from the large subunit of *T. thermophilus* and cut the region previously identified as the PTC. The picture was edited by hand using image editors. Plus, a predicted secondary structure for the PTC of *T. thermophilus* (comprising 179 bp) was generated using the software RNAstructure [34] and visualized in the foRNA applet [35]. Both structures were colored according to the variation of information of the PTC alignment of all the 1425 sequences. Both the sequence and the 3D structure of *T. thermophilus* 23S rRNA (PDB ID 4v4i) were downloaded and manually edited to obtain the PTC region only.

### 2.6. 3D Modeling of the Different PTC Sequences and Structural Comparisons

Using the UGENE software [36], consensus sequences were obtained from each alignment file. The ModeRNA webserver [37] was used to perform template-based 3D modeling using the *T. thermophilus* PTC as a model to predict the 3D structure of the consensus sequences from the three datasets under analysis. The modeled structures were structurally aligned using the RNA-align software from the Zhang Lab suite [38]. Finally, we used Chimera [39] to visualize the structural comparisons.

## 3. Results

### 3.1. PTC Datasets: Production and Taxonomic Analysis

On 7 October 2019, the GenBank database contained 1434 complete sequences of 23S ribosomal RNAs. All these sequences were downloaded in FASTA format and aligned (using an optimal pairwise alignment tool) to the PTC of the bacteria *T. thermophilus* to map a PTC sequence region containing 179 bp. Using in-house Perl scripts, we parsed the PTC alignment information and generated a file containing 1434 PTC sequences. After an initial round of multiple alignment of the PTC dataset, we visually identified ten sequences that seemed too divergent in the alignments. These anomalous PTC sequences were removed from further analyses as they possibly represented sequences with inaccurate genome annotation [40]. Therefore, a dataset containing 1424 PTC sequences in FASTA format was produced. This dataset presented five sequences from eukaryotes, 118 sequences from archaea and 1301 sequences from bacteria. Inside the bacterial clade, we observed that the major groups sampled were Proteobacteria (564 sequences), Firmicutes (237 sequences), and Actinobacteria (153 sequences); another 347 sequences were divided into 25 other clades. From the Archaea domain, 81 sequences came from Euarchaeota, 33 from Crenarchaeota, three from Thaumarchaeota, and one from Nanoarchaeota. The five sequences from eukaryotes came from the fungi species *Encephalitozoon intestinalis*. Table 1 summarizes the information about the sequences obtained.

**Table 1.** Peptidyl transferase center (PTC) sequences per clade, number of sequences, and multiple alignment features.

Dataset Name	Clades	Number of PTC Sequences	Alignment Size (Gaps)	Positions 100% Conserved
PTC-all	Bacteria	1424	201 (10)	42
	Archaea			
	Eukarya			
PTC-Bac	Bacteria *	1301	195 (8)	62
PTC-Arc	Archaea *	118	186 (7)	83
PTC-Pro	Proteobacteria *	564	186 (7)	110
PTC-Fir	Firmicutes **	237	184 (4)	122
PTC-Act	Actinobacteria	153	179 (0)	132

\* These datasets are subsets of the PTC-all dataset. \*\* These datasets are subsets of the PTC-Bac dataset.

### 3.2. Multiple Alignment and Sequence Conservation of PTC Datasets

A multiple alignment of each PTC dataset was performed and followed by the analysis of sequence conservation in the main datasets. Table 1 shows that all dataset alignments were 179 bp in length plus the number of gaps added by the alignment tool to optimize the sequence alignment. As expected, due to the presence of highly divergent sequences, the dataset PTC-all presented the highest number of gaps (10) and the fewest number of 100% conserved nucleotides (42). The highest number of positions 100% conserved was found in the Actinobacteria dataset, with 73.7% of conserved nucleotides and no gaps found. This was followed by Firmicutes, with 66.3% of residues completely conserved, and Proteobacteria, with 59.1%, evidencing a higher diversity of PTC sequences in the latter clade. This possibly happened due to the existence of large sub-clades (such as Alpha-, Beta, Gamma-, and Deltaproteobacteria), presenting well-differentiated PTC sequences among them.

The multiple alignments of the three main PTC datasets, i.e., PTC-all, PTC-Bac, and PTC-Arc, are displayed as sequence logos (Figure 1), on which the conservation profile can be easily visualized. In sequence logos, the 100% conserved positions were shaded in green, and the main adenine located at the catalytic site was starred on top. The observed single nucleotide gaps reveal variability in size among sequences. In some cases, the difference was detected in a single nucleotide present in one individual sequence, such as (i) the gap observed in position 21 of the PTC-all dataset; (ii) position 169 in PTC-all (and its equivalent position 165 in PTC-Bac) appeared merely due to the presence of a G in the bacterium *Spirochaeta africana*; and (iii) position 16 in PTC-Bac that was represented due to a C observed solely in the bacterium *Streptococcus pyogenes*. Other exceptional cases of gaps include positions 106 and 120 in PTC-all (equivalent to positions 101 and 115 in PTC-Bac) that correspond to single nucleotides observed in two Gammaproteobacteria species. In other cases, the differences rely on a few sequences that belong to one specific clade; for example, the gap in position 53 in PTC-all and position 49 in PTC-Bac correspond to a T present in Firmicutes species.

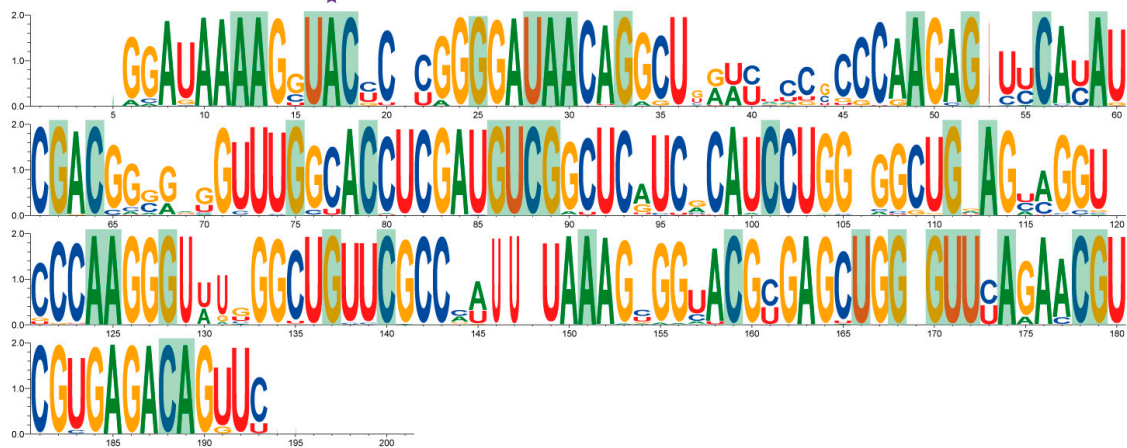
The lowest nucleotide conservation observed in the PTC sequences spans from nucleotide 37 to 44 and 65 to 70 in PTC-all and happened due to the fact that several bacteria from the phyla Chloroflexi, Aquificae, Fusobacteria, Bacteroidetes, Thermotogae, and Elusimicrobia (as well as some archaea) present divergent nucleotides in that region. In PTC-Bac, the sections ranging from 32 to 41 and 62 to 66 are observed to be more variable because the sequence data for the aforementioned phyla do not have analogous counterparts in archaea.

Considering the gaps spanning more than one nucleotide, it is possible to see a void at the beginning of the PTC-all alignment. It corresponds to a certain variation observed in the 5' PTC region of some euryarchaeal sequences that seem to present a duplication in their five initial nucleotides. Such a region has a more notorious representation in the PTC-Arc dataset. The gap observed at the 3' end of the PTC-all alignments and nearly at the end of the PTC-Bac alignments appeared due to one single sequence coming from the cyanobacterium *Thermosynechococcus elongatus* that differs from all other organisms. Therefore, the multiple alignment tool positioned the sequences in the most convenient way, either keeping the gap at the very end of the PTC-all dataset alignment (from nucleotides 194 to 201) or introducing a gap just before the end, as observed in PTC-Bac (between sites 181 and 185).

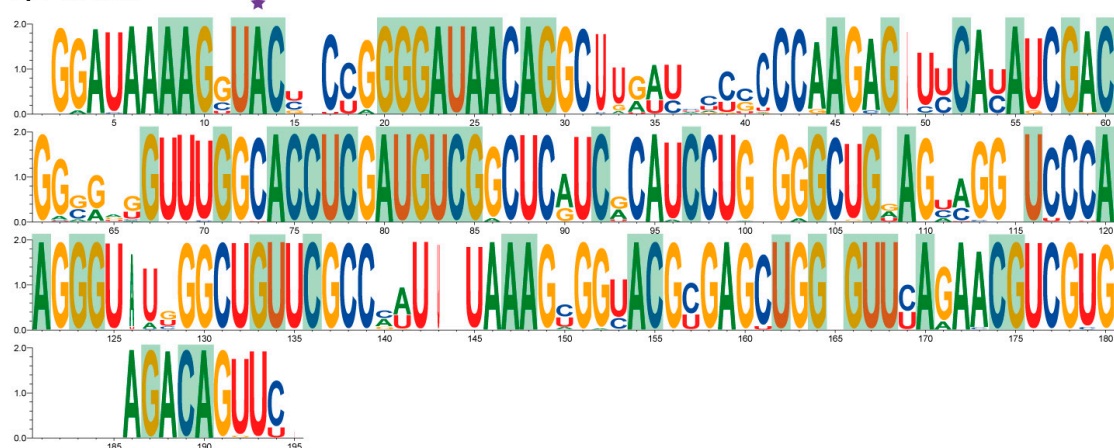
Interestingly, the only gaps observed in the PTC-Arc dataset appear at both ends (5' and 3'). The initial one corresponds to the previously mentioned apparent duplication of the initial nucleotides in some euryarchaeal sequences. There is also a short portion at the last two nucleotides of the PTC-Arc dataset that shape an apparent gap at the end of alignment. This is observed because the PTC from *Nanoarchaeum equitans* and from some crenarchaeal sequences have a couple of G nucleotides inserted there. Additionally, the PTC-Arc sequence logo (Figure 1c) shows the highest amount of heterogeneity among all three datasets, observed in sites with different sizes of nucleotides along the vertical axis. Notably, even if the PTC-Arc dataset presented the highest number of conserved sites along the two domains analyzed (83 sites shaded in green for archaea compared to 62 for bacteria), its non-conserved

sites also displayed higher variation. This fact denotes both the conservation of a tight PTC structure and a significant variability of this diverse clade that originated eukaryotic organisms [27,28].

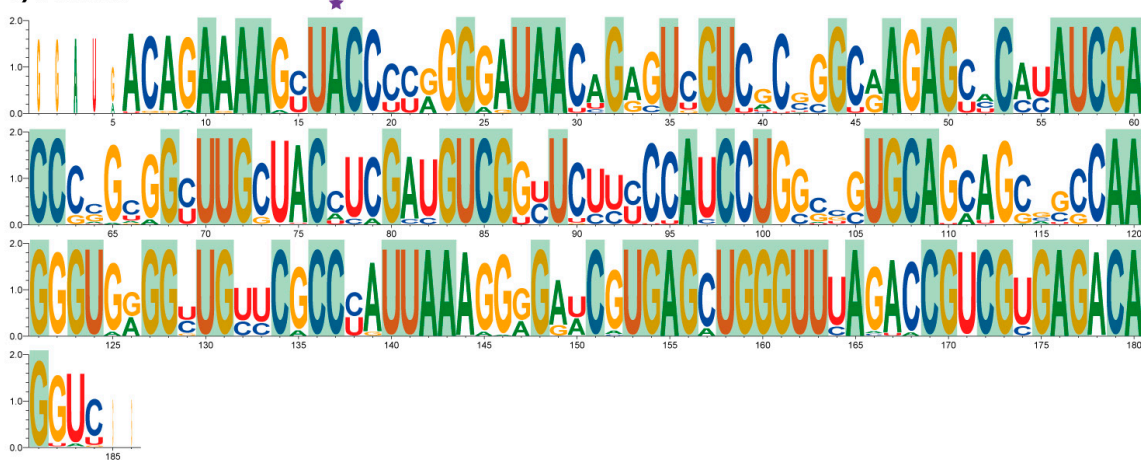
### a) PTC-all



### b) PTC-Bac



### c) PTC-Arc

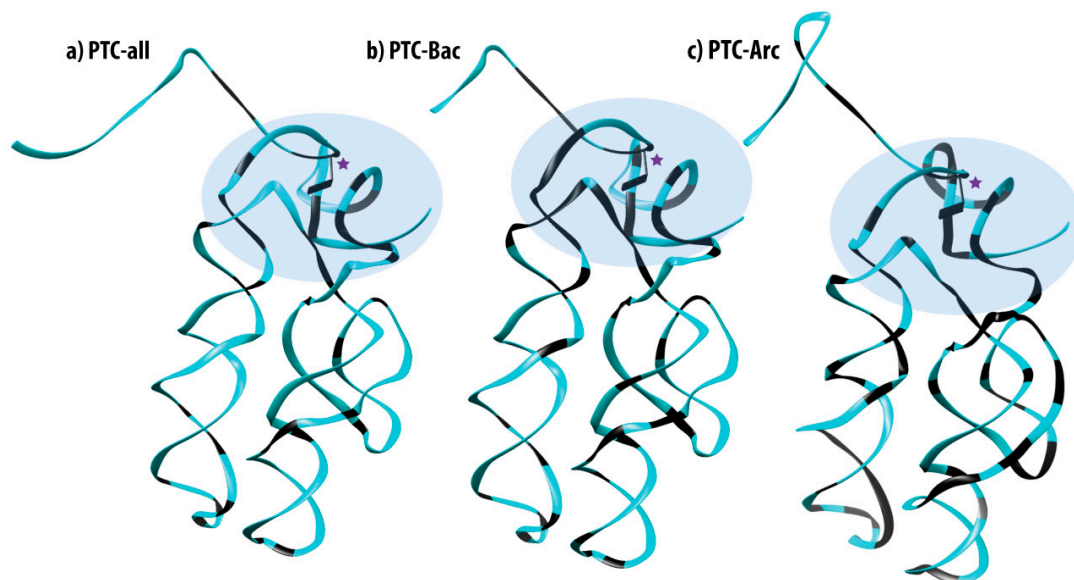


**Figure 1.** WebLogos showing nucleotide conservation in the main analyzed PTC datasets. (a) PTC-all; (b) PTC-Bac; (c) PTC-Arc. Universally conserved nucleotides in each dataset are shown with a green background. The adenine located at the catalytic site is highlighted with a magenta star.

Finally, site A2451 from the 23S rRNA is the catalytic site of the PTC, essential for the peptide bond to occur. This nucleotide has been shown to be absolutely preserved in each and every analyzed sequence, as it can be observed at the highlighted site with a magenta star at A17 in PTC-all, A13 in PTC-Bac, and A17 in PTC-Arc (Figure 1).

### 3.3. Mapping 100% Conserved Sites into the 3D Structure of the PTC

To gain insights about the nucleotide conservation in the PTC, we produced 3D models in which the 100% conserved positions (shaded green in Figure 1 and drawn in black in Figure 2) could be seen over the tridimensional structure. Thus, we obtained the PTC consensus sequences for each dataset and modeled their 3D structures using ModeRNA software using the known PTC structure for *T. thermophilus* (PDBid 4V4I) as a template. Analyzing Figure 2, we observe a higher number of conserved positions aggregated at the top of the structure (shaded oval in Figure 2), close to the catalytic site A2451 (identified with a magenta star). Nevertheless, the entire structure is significantly conserved, and specific conserved nucleotides located at different sites along the whole structure are probably anchors for holding the 3D shape of the PTC.



**Figure 2.** Embedding the 100% conserved nucleotides for each dataset: (a) PTC-all, (b) PTC-Bac, and (c) PTC-Arc. The structures were produced by template-based modeling over the known *T. thermophilus* structure. The catalytic site A2451 is depicted as a rectangle protrusion close to a magenta star. The 100% conserved nucleotides are colored in black and ovals accentuate the regions with highly conserved sites at the top of the structures.

### 3.4. Identity Elements, Entropy Variation, Information Variation

For the complete set of sequence alignments, an entropy value was determined for each position. Thus, nucleotides were grouped into information clusters that were identified by color codes along the sequences (Figure S1 in Supplementary). The first graph (Figure S1a) corresponds to the analysis of both bacterial and archaeal sequence alignments, while Figure S1b,c present the data for bacterial and archaeal sequences, respectively. Given that the number of sequences used for the informational analysis decreased between the PTC-all dataset and the archaeal one, the number of information clusters and positions in the clusters increased due to the reduction in variation. Ungapped positions, such as 45, 73, and 116 observed in the PTC-all dataset, have shown entropy close to the maximum value of two, meaning that all four nucleotides occurred in almost equiprobable amounts. Colored clusters with entropy equal to zero represent invariant nucleotide positions (Figure S1); while clusters with entropy greater than zero reflect positions on which nucleotide variations are highly predictable.

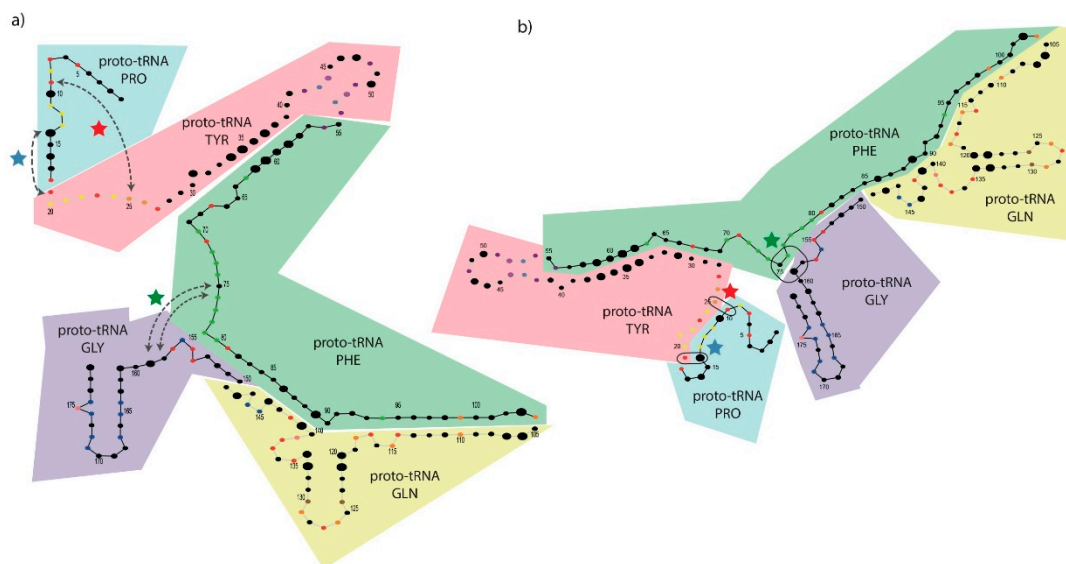


This property can be clearly noticeable in the entropy profile for archaeal sequences (Figure S1c) in which the positions in the red-colored cluster have entropy greater than 0. In each plot of Figure S1, the red color is associated with the modal cluster, i.e., the cluster containing the highest number of interdependent, co-evolving nucleotide positions.

The PTC-all dataset showed 25 nucleotide positions grouped into nine information clusters. Six of these clusters contained two positions and the remaining three clusters contained six, four, and three positions each. The information cluster harboring three positions was the only one in PTC-all whose bases were invariant, i.e., showed entropy equal to zero. The PTC-Bac dataset presented 71 positions divided into 11 clusters: Five clusters contained two positions, whereas the others presented 20, 11, 10, 8, 7, and 5 nucleotide positions. The cluster with the highest number of positions (20) was the only one that presented an intra-cluster entropy of zero. Regarding the PTC-Arc dataset, 101 positions were found split into 16 information clusters: 12 clusters contained two positions and the others possessed 47, 20, 7, and 3 positions. In the archaea data, the cluster containing 20 nucleotides was unique and contained an intra-cluster entropy equal to zero, representing invariant positions. All the nucleotide positions of the clusters are shown in Supplementary Table S1.

### 3.5. Mapping Identity Elements and Information Clusters into the 2D Structure of the PTC

A subtle variation was produced for each dataset alignment to consider only their sequence alignments without gaps, so that the length matched the canonical 179 nucleotides from the PTC. Even knowing that gaps are key, and their removal might produce artefacts, we proceeded to choose fixed length alignments to try to better explore the results, as the following analyses required these sorts of input data. Thus, information clusters were computed one more time and mapped into both (i) the known secondary structure of the PTC from *T. thermophilus* (Figure 3a) and (ii) the secondary structure predicted de novo by the software foRNA, using the PTC-all consensus sequence as entry (Figure 3b).



**Figure 3.** Information clusters and proto-tRNA composition from the PTC-Bac dataset as observed over (a) the secondary structure of the modern PTC from *T. thermophilus* and (b) the de novo (predicted) RNA structure. The radius of each circle corresponds to its entropy value, i.e., bigger circles represent more variable positions. Regions corresponding to each proto-tRNA are shown in colored boxes according to the corresponding ancestors. Colored stars in blue, red, and green represent Watson–Crick base pairing in putative ancestral folding (b) that were separated to generate the modern folding (a). Arrows in (a) represent the positions linked by Watson–Crick base pairing in (b) that were separated to produce the catalytic structure of the PTC (a).

Analyzing Figure 3, one observes that the de novo structure differs from the modern PTC in two significant regions: (i) A stem formed on the first segment by the pairings of the bases from 9:25 (Figure 3, red star) up to 14:19 (Figure 3, blue star) and forming a loop with the segment 15–18 (Figure 3b) and (ii) the extra pairings of bases 75:159 and 76:158 (Figure 3, green star), on which the segment that joins the two coils in the secondary structure of the PTC can be seen. These three stars represent the main topological differences in the 2D structures (shown in Figure 3a,b), as all the other nucleotides are arranged similarly. The stars indicate which Watson–Crick base pairings should be broken in the lower entropy arrangement of Figure 3b to produce a modern-like PTC structure, as observed in Figure 3a. It is possible that these two RNA structures were viable and interchangeable at the origins of the PTC in prebiotic Earth, as it is known that RNA molecules can adopt different conformations [41–43]. These alternative foldings (among other possible ones) possibly changed according to the presence of different ligands and environmental conditions [44].

### 3.6. Mapping Proto-tRNAs into the 2D Structure of the PTC

When trying to explain how the PTC might have been formed in the past, we benefited from the work of Farias [45], who obtained putative ancestral tRNAs (Supplementary Information) using a dataset of 9758 sequences downloaded from a tRNA database [46]. In that work, Farias (2013) [45] separated 22 types of tRNAs, including 20 canonical tRNAs, one initiator tRNA, and one tRNA for selenocysteine, ran ModelTest to find the best nucleotide substitution model, and produced ancestral sequences using an approach based on maximum likelihood. In a following publication, Farias and collaborators (2014) [19] used a combinatorics approach to randomly concatenate those ancestral proto-tRNAs and search for possible matches in a nucleotide alignment against protein databases. Notably, these researchers found a specific combination of five proto-tRNAs concatenated directly (+/+ strands) that was shown to present 50% nucleotide identity to the PTC of the bacterium *T. thermophilus*. Therefore, to check whether the early origin of the PTC could be explained by the concatenation of those proto-tRNAs, we took these ancestral proto-tRNAs that bound to the amino acids (i) proline (Pro), (ii) tyrosine (Tyr), (iii) phenylalanine (Phe), (iv) glutamine (Gln), and (v) glycine (Gly) and aligned them to the PTC of *T. thermophilus*. These five proto-tRNAs were therefore mapped (in the order described above) within the segments 1–18, 19–54, 55–104, 105–149, and 150–179 of the modern PTC from *T. thermophilus* and plotted in colored boxes, as illustrated in Figure 3. As observed in the original publication (Farias, 2013) [45], the ancestral proto-tRNAs presented variable sizes (as measured in base pairs) due to the variable nucleotide conservation observed in the modern tRNAs used to produce them. Therefore, the maximum likelihood model, applied by Farias (2013) [45], removed some nucleotide positions that were not shown to be conserved in most tRNAs used to build the ancestral sequences and produced a sort of “truncated” ancestral proto-tRNA. The higher the nucleotide conservation in the modern tRNA sequences (to build the ancestral sequences), the longer the proto-tRNAs were.

We proceeded to analyze the co-occurrence of those proto-tRNAs and information clusters to check whether it could provide us with some insights. To do that, we started analyzing the PTC-Bac dataset due to the fact that it presented an intermediate number of clusters, as PTC-all contained too few positions in clusters (25 nucleotides), and PTC-Arc presented too many (101 nucleotides). Thus, analyzing the bacterial dataset (containing 71 nucleotides in 11 clusters), we noticed that four out of nine information clusters contained bases corresponding to the positions located in nucleotides placed in regions mapped in two different proto-tRNAs. In particular, Figure 3 provides evidence that (i) the cluster colored in yellow contained bases (black circles) putatively coming from proto-tRNA<sup>Pro</sup> and proto-tRNA<sup>Tyr</sup>; (ii) the purple cluster contained bases (black circles) within proto-tRNA<sup>Tyr</sup> and proto-tRNA<sup>Phe</sup>; (iii) the orange cluster contained bases (black circles) found in proto-tRNA<sup>Phe</sup> and proto-tRNA<sup>Gln</sup>; and (iv) the dark blue cluster contained bases (black circles) found in both proto-tRNA<sup>Gln</sup> and proto-tRNA<sup>Gly</sup>. We hypothesize that these information clusters represent co-evolving nucleotides originally responsible for linking the proto-tRNAs together in a higher-level

secondary structure of extreme relevance to shaping the overall PTC structure. The cluster represented by (v), shown in red dots, contained bases mapped in all proto-tRNAs and was possibly relevant to the assembly of the whole 3D structure of the PTC. The one (vi) cluster shown with green dots contained a segment corresponding only to the third proto-tRNA<sup>Phe</sup> (Figure 3). The other five clusters from the PTC-Bac dataset contained merely two bases representing Watson–Crick base pairs inside regions mapping single proto-tRNAs.

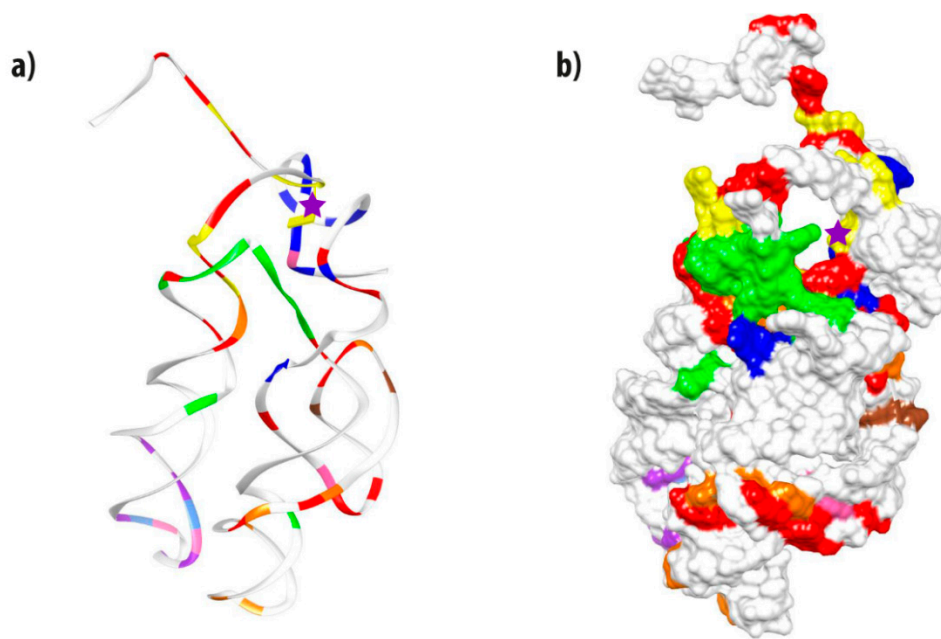
Similar data about the relationship between proto-tRNA mapping and information clusters are presented for PTC-all and PTC-Arc (Figure S2; Supplementary Table S1). Regarding the PTC-all dataset, we found one cluster containing six positions to have nucleotides coming from all the five proto-tRNAs. This finding reinforces the previous hypothesis of PTC formation by the concatenation of proto-tRNAs and suggests that this informational relationship might help to bind the proto-tRNAs together. Another identity cluster containing four positions was found in sites present in both proto-tRNA<sup>Gln</sup> and proto-tRNA<sup>Gly</sup>. The PTC-all cluster, with three positions, embraced the first, fourth, and fifth proto-tRNAs and two clusters with two bases were found to present nucleotides in two regions mapped to different proto-tRNAs. Regarding the PTC-Arc dataset, the two clusters containing the highest number of nucleotide positions (47 and 20), encompass regions mapped in all the five proto-tRNAs. The PTC-Arc cluster, containing seven nucleotides, mapped into the third, fourth, and fifth proto-tRNAs. A cluster with three positions was mapped into the first, fourth, and fifth proto-tRNAs. Finally, six out of 12 clusters, containing two positions, mapped to two different proto-tRNAs. It is also conspicuous that in both the PTC-all and PTC-Arc datasets, there exist identity clusters containing nucleotide positions shared by non-consecutive proto-tRNAs. Altogether, those mappings reinforce the hypothesis that these information clusters account for the 3D configuration of the modern PTC by linking together ancestral proto-tRNAs.

### 3.7. Mapping Identity Elements and Information Clusters into the 3D Structure of the PTC

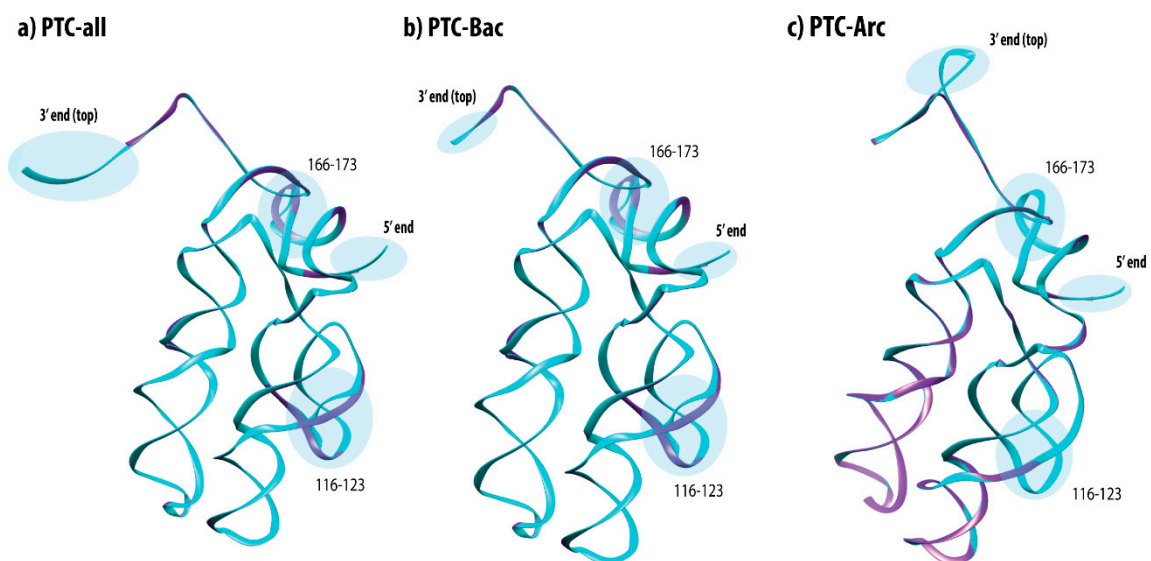
In Figure 4, the tridimensional structure of the PTC derived from *T. thermophilus* 23S rRNA is shown and colored according to the different information clusters found in the PTC-Bac dataset, aiming to observe their spatial distribution. For all datasets, the most evident characteristic is that many identity clusters bundle in a tridimensional configuration (data for PTC-all and PTC-Arc are shown in S3). Besides the modal cluster (in red), most other nucleotides sharing similar clusters are observed in nearby positions. Brown sites, for example, remain together in the arm observed at the right side of the structure in Figure 4. Most of the green and dark blue nucleotides are close together in the 3D shape, near the yellow cluster that contains the catalytic site (A12 in this structure, marked with a magenta star). The red-colored nucleotides—corresponding to the cluster containing a higher number of interrelated, co-evolving positions—are spread throughout the PTC molecule, possibly responsible for maintaining the whole structure. They are absent, however, from the bottom of one arm (observed at the left side of Figure 4a) in which purple, light blue, and pink clusters mostly muster. Therefore, all clusters seem necessary to maintain the PTC structure and create the cage necessary for the peptide bonds to occur.

### 3.8. Structural Alignment of PTC Datasets to *T. Thermophilus*

The template-based structural reconstructions for the consensus sequences from the PTC-all, PTC-Bac, and PTC-Arc datasets were structurally compared to the actual PTC known for *T. thermophilus* in order to provide an appreciation of the spatial differences among them (Figure 5; in which cyan-colored regions represent a match between predicted and actual PTC structures and purple represents differences).



**Figure 4.** Ribbon (a) and surface (b) tridimensional representations of the PTC from *T. thermophilus* with the information clusters found in the PTC-Bac dataset colored according to Figure 3. The catalytic site (corresponding to A2451 or A12 in the current model) is marked with a magenta star and protrudes from the ribbon sketch (a).



**Figure 5.** Tridimensional structural comparisons between the template PTC structures from *T. thermophilus* are shown in purple and the predicted PTCs, depicted in cyan for the different clades: (a) PTC-all, (b) PTC-Bac, and (c) PTC-Arc. Relevant regions are encircled and labeled.

As expected, due to the high general conservation shared among their sequences, the template-based structures were similar to the PTC of *T. thermophilus* [47], while the most relevant variations were observed towards their extremities. Indeed, the most similar structure to *T. thermophilus*' PTC was PTC-Bac, which was shown to be only slightly longer than the model at the 3' end of the structure. The most patent difference between the model and the consensus of PTC-all was the presumptive insertion of five nucleotides at the 3' end, whereas it was arranged as an internal insertion positioned at the top of the PTC-Arc structure.

In the portions corresponding to nucleotides 116 to 123 and 166 to 173 of the PTC from *T. thermophilus*, both PTC-all and PTC-Bac consensuses differed from the model. However, the alignments revealed only a slight difference in the corresponding segments. An opposite situation happened in the comparison to the PTC-Arc structure, in which the sites 116 to 119 revealed a highly variable sequence region though the consensus structure and the models were shown to be nearly undistinguishable.

#### 4. Discussion

The sequence of the PTC is possibly the most relevant stretch of nucleic acid to be studied if one aims to understand the origin of life. Nowadays, it is a consensus that the ribosome should be understood as a prebiotic machine that predated the origin of cells [16,48]. Although there is a debate in the literature about whether the ribosome was built over the PTC region or not [9–11], many researchers claim to have evidence that the assembly of the genetic code and the ribosome started with the initial formation of region V of the ribosome, just the place in which the PTC is settled. Theoretical works on the origin of life suggest that the contingent appearance of this ribozyme capable of binding amino acids together was crucial to both the initial emergence and further development of the phenomenon of life [7,49,50].

Here, we evaluated the sequence and structural conservation of the peptidyl transferase center using all completely available sequences of 23S ribosomal RNA present in the GenBank database and annotated as such. We decided to use complete sequences to add rigor to the analyses and to avoid to the sequencing errors often present in small molecules [40]. Besides, as we were interested in understanding the relevance of the PTC to the early origin of life, we decided to exclude eukaryotic sequences from the analyses. Eukaryotes are now known to have originated from archaeal organisms coming from the phylum Lokiarchaeota, subphylum Asgard [27,28], therefore being derivate clades and having no substantial role in early origins of life.

The 1424 23S rRNA sequences obtained were aligned, filtered out to retrieve only the PTC region, and divided into three main datasets. Although there is a consensus in the literature about the fact that the PTC sequence is highly conserved [51–55], to our knowledge, PTC sequences have not been analyzed thoroughly by comparative sequence analysis, information variation, and bi/tri-dimensional structural analysis to better validate this assumption.

The multiple alignment comparison showed that the PTC from archaea presents about 40% fully conserved bases; this number lowers to 30% in bacteria, and to 20% in all analyzed organisms (Table 1). These percentages of conservation are clearly dependent on the total number of sequences analyzed, as bacteria possessed >11x more sequences available in GenBank than archaea (1302 versus 118). Curiously, the archaeal dataset presented both a higher number of conserved nucleotides (83) and a higher number of variable sites when compared to the bacterial one. This possibly means that the structure of the archaeal PTC is more optimized to be tighter in specific regions that maintain a rigid tridimensional backbone and looser in others. By contrast, the structure of the bacterial PTC is possibly wobblier.

When observed in the context of the bi- and tridimensional structures, we found that most fully conserved bases from the PTC folded close to the catalytic site, whereas sites located down to the two hairpin structures seem to allow more variation (Figure 2). This last fact should be expected, as the catalytic sites of enzymes are often more conserved than the other parts; the same is true for ribozymes. The PTC is known to be a flexible and efficient catalyst [12] as it is capable of recognizing different, specific substrates (20 different amino acids bind to aminoacyl-tRNAs) and polymerizing proteins at a similar rate [56,57]. Therefore, considering the extreme relevance of the PTC, it would be surprising if the site of catalysis showed variation.

The use of a pseudometric to show the information variation in PTC sequences allowed us to identify clusters of nucleotides that are informationally linked. We were able to find clusters containing as many as 47 nucleotides, although most of them presented fewer than 10 nucleotides. The clusters

with a higher number of bases were colored in red for all datasets and they were invariant for PTC-all and PTC-Bac, although in PTC-Arc, this modal cluster presented 47 nucleotide positions that could vary coordinately. We hypothesize that these clusters were mainly important to keep the tridimensional structure of the PTC, but we found out that they also provided interesting insights about how the PTC was formed.

Farias and collaborators [19] used tRNA sequences from hundreds of species, together with maximum likelihood analyses, to construct ancestral sequences for each of the 20 different tRNAs, producing putative ancestral proto-tRNAs. When they randomly concatenated these proto-tRNAs and BLASTed them with GenBank's nucleotide database, they verified that one concatemer of five proto-tRNAs presented a significant nucleotide identity (about 50%) to the 23S rRNA of the bacterium *T. thermophilus*, exactly in the PTC region [19]. Even if 50% identity cannot be considered a significant threshold for sequence identity, one cannot expect to apply modern standard measures of nucleotidic variation when working with an event so distant in the past. Therefore, even considering the hypothetical nature of this result, we decided to go further into that investigation. Thus, we mapped their five proto-tRNAs concatamers into the 2D structure of the *T. thermophilus* PTC. Besides mapping the proto-tRNAs, we also produced a diagram in which the nucleotides were colored according to the clusters produced with the information variation analysis. This resulted in Figure 3a, which showed the 2D structure of the PTC with dots representing each nucleotide of the PTC colored according to its corresponding cluster. We found out that many information clusters contained informationally linked nucleotides mapped to distinct proto-tRNAs along the PTC structure. This fact seems to indicate that these nucleotides may have been relevant to the stepwise binding of these ancestral tRNAs with each other in order to produce the modern shape of the PTC site. These results are in accordance with recent works suggesting that either the PTC or rRNAs should have been formed by the assembly of tRNAs [17,18,20,22]. We not only confirmed these previous assumptions but added new information on the conserved sites possibly used to link the PTC structure. Additionally, we used a de novo modeling software to predict the 2D structure of the PTC (Figure 3b) and were able to produce a structure very similar to the modern PTC. Working over this structure, we were able to identify four sites linked by Watson–Crick bonds that, when released, may have given rise to the modern PTC (these sites were identified by three colored stars in Figure 3). Our hypothesis is that the ancient PTC could be observed in at least these two structures, changing from one to the other according to the presence of ligands and specific environmental conditions—a known property of RNAs—to present multiple unstable, interchangeable structures [41,44]. Additionally, a hairpin with yellow dots observed in the bottom part of the de novo structure (Figure 3b) clearly indicated which nucleotides were used to bind proto-tRNA<sup>Pro</sup> and proto-tRNA<sup>Tyr</sup> into an integrated, higher-level structure that produced the PTC. Although the binding of other nucleotides between distinct proto-tRNAs cannot be clearly observed in the current structure of the PTC, we hypothesize that these co-evolving nucleotides were important to bind the proto-tRNAs together when the PTC was under formation, as its secondary structure probably grew by the addition of one tRNA at time. Thus, the informationally linked nucleotides possibly held the higher-level 2D and 3D structures together to allow the stepwise formation of the whole PTC region.

As both the position of nucleotides sharing the same cluster (Figure S1) and the bidimensional structure of the PTC (Figure 3) looked sometimes distant and peculiar, we decided to gain new insights by plotting the clusters into a tridimensional structure. We used the same color code for clusters to check whether the position of the nucleotides sharing same clusters would make more sense in 3D and we found that this was indeed the case for most clusters (Figure 4). Both the ribbon and the surface structures demonstrate that nucleotides sharing the same information cluster were usually observed close to each other in the 3D structure.

An interesting possibility derived from the current analyses would be the actual production of resurrection experiments [58] able to synthesize the putative form of an ancient PTC using the exact nucleotide sequence derived from the concatemer of proto-tRNAs described here. Which properties

may this molecule have? How would it fold? Could this sequence function as a ribozyme and catalyze a peptide bond? Similarly, it could be possible to synthesize the proto-tRNAs proposed by Farias (2013) [45] and verify whether they could bind with each other using the nucleotides described in the information variation model we used. Those experiments could bring an experimental background to the theoretical analyses performed here.

Finally, the predicted 3D structures based in the consensus sequences for each dataset were structurally compared to the known PTC structure from *T. thermophilus* [47] to allow for the identification of similarities and divergences. Despite some notable nucleotide variations from PTC-all and PTC-Bac to the *T. thermophilus* sequence, the 3D structure was shown to be significantly conserved. The higher divergences found were to be related to slight extensions observed in the 3' regions of the datasets. Regions containing nucleotide variance were shown to be conserved at the structural level (Figure 5). In conclusion, we have provided (i) a better understanding of how nucleotide variation is observed in the PTC, underscoring (ii) a testable possible model about how proto-tRNAs shaped its structure, and (iii) how the evolutionary process froze essential nucleotide positions that enabled the peptide polymerization bonding by preserving the tridimensional structure of the PTC ribozyme.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2075-1729/10/8/134/s1>, Figure S1: Variation of information (bits) along the sequence of the PTC: (a) PTC-all (b) PTC-Bacteria, and (c) PTC-Archaea. Figure S2: Relationship between proto-tRNA mapping and information clusters for (a) PTC-all and (b) PTC-Arc; Figure S3: Mapping identity elements and information clusters to the 3D structure of (a) PTC-all (b) PTC-Bacteria, and (c) PTC-Archaea. Table S1: Identification of nucleotide positions clustered by the information analysis of each PTC dataset. The tRNA ancestral sequences.

**Author Contributions:** Conceptualization: F.P., M.V.J., S.T.d.F.; methodology: F.P., G.S.Z., M.P.-P., M.V.J.; software F.P., G.S.Z., M.P.-P.; validation: F.P., G.S.Z., M.P.-P., S.T.d.F.; formal analysis: F.P., G.S.Z., M.P.-P., S.T.d.F., M.V.J.; investigation: F.P., G.S.Z., M.P.-P., S.T.d.F., M.V.J.; resources M.V.J.; data curation: F.P., G.S.Z., M.P.-P.; writing—original draft preparation F.P., S.T.d.F., M.V.J.; writing—review and editing: F.P. and M.V.J.; visualization F.P., G.S.Z., M.P.-P.; supervision F.P., S.T.d.F., M.V.J.; project administration: F.P. and M.V.J.; funding acquisition M.V.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** We thank FAPERJ (CNE E-26/202.780/2018) and CNPq (PDE 205072/2018-6) for funding FP. GSZ and MPP are doctoral students from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM), and receive doctoral fellowships from CONACYT, numbers 737920 and 694877, respectively. MVJ was funded by Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México, UNAM (PAPIIT-IN201019).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Noller, H.F.; Woese, C.R. Secondary structure of 16S ribosomal RNA. *Science* **1981**, *212*, 403–411. [[CrossRef](#)]
2. Noller, H.F.; Hoffarth, V.; Zimniak, L. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **1992**, *256*, 1416–1419. [[CrossRef](#)]
3. Lohse, P.A.; Szostak, J.W. Ribozyme-catalysed amino-acid transfer reactions. *Nature* **1996**, *381*, 442–444. [[CrossRef](#)] [[PubMed](#)]
4. Lanier, K.A.; Petrov, A.S.; Williams, L.D. The Central Symbiosis of Molecular Biology: Molecules in Mutualism. *J. Mol. Evol.* **2017**, *85*, 8–13. [[CrossRef](#)] [[PubMed](#)]
5. Vitas, M.; Dobovišek, A. In the Beginning was a Mutualism—On the Origin of Translation. *Orig. Life Evol. Biosph.* **2018**, *48*, 223–243. [[CrossRef](#)] [[PubMed](#)]
6. De Farias, S.T.; Prosdocimi, F. *A Emergência dos Sistemas Biológicos: Uma Visão Molecular da Origem da Vida*, 1st ed.; ArtcomCiencia: Rio de Janeiro, Brazil, 2019; ISBN 978-65-900624-1-3.
7. Prosdocimi, F.; José, M.V.; de Farias, S.T. The First Universal Common Ancestor (FUCA) as the Earliest Ancestor of LUCA's (Last UCA) Lineage. In *Evolution, Origin of Life, Concepts and Methods*; Pontarotti, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 43–54. ISBN 978-3-030-30363-1.
8. Bokov, K.; Steinberg, S.V. A hierarchical model for evolution of 23S ribosomal RNA. *Nature* **2009**, *457*, 977–980. [[CrossRef](#)]

9. Petrov, A.S.; Bernier, C.R.; Hsiao, C.; Norris, A.M.; Kovacs, N.A.; Waterbury, C.C.; Stepanov, V.G.; Harvey, S.C.; Fox, G.E.; Wartell, R.M.; et al. Evolution of the ribosome at atomic resolution. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 10251–10256. [[CrossRef](#)]
10. Caetano-Anollés, G. Ancestral Insertions and Expansions of rRNA do not Support an Origin of the Ribosome in Its Peptidyl Transferase Center. *J. Mol. Evol.* **2015**, *80*, 162–165. [[CrossRef](#)]
11. Petrov, A.S.; Williams, L.D. The ancient heart of the ribosomal large subunit: A response to Caetano-Anollés. *J. Mol. Evol.* **2015**, *80*, 166–170. [[CrossRef](#)]
12. Rodnina, M.V. The ribosome as a versatile catalyst: Reactions at the peptidyl transferase center. *Curr. Opin. Struct. Biol.* **2013**, *23*, 595–602. [[CrossRef](#)]
13. Caetano-Anollés, G.; Caetano-Anollés, D. Computing the origin and evolution of the ribosome from its structure—Uncovering processes of macromolecular accretion benefiting synthetic biology. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 427–447. [[CrossRef](#)] [[PubMed](#)]
14. Lehmann, J. Induced fit of the peptidyl-transferase center of the ribosome and conformational freedom of the esterified amino acids. *RNA* **2016**, *23*, 229–239. [[CrossRef](#)] [[PubMed](#)]
15. Agmon, I.; Bashan, A.; Zarivach, R.; Yonath, A. Symmetry at the active site of the ribosome: Structural and functional implications. *Biol. Chem.* **2005**, *386*, 833–844. [[CrossRef](#)] [[PubMed](#)]
16. Belousoff, M.J.; Davidovich, C.; Zimmerman, E.; Caspi, Y.; Wekselman, I.; Rozenszajn, L.; Shapira, T.; Sade-Falk, O.; Taha, L.; Bashan, A.; et al. Ancient machinery embedded in the contemporary ribosome. *Biochem. Soc. Trans.* **2010**, *38*, 422–427. [[CrossRef](#)]
17. Tamura, K. Ribosome evolution: Emergence of peptide synthesis machinery. *J. Biosci.* **2011**, *36*, 921–928. [[CrossRef](#)]
18. Caetano-Anollés, G.; Sun, F.-J. The natural history of transfer RNA and its interactions with the ribosome. *Front. Genet.* **2014**, *5*, 5.
19. De Farias, S.T.; Rêgo, T.G.; José, M.V. Origin and evolution of the Peptidyl Transferase Center from proto-tRNAs. *FEBS Open Bio* **2014**, *4*, 175–178. [[CrossRef](#)]
20. Root-Bernstein, M.; Root-Bernstein, R. The ribosome as a missing link in the evolution of life. *J. Theor. Biol.* **2015**, *367*, 130–158. [[CrossRef](#)]
21. De Farias, S.T.; Rêgo, T.G.; José, M.V. Peptidyl Transferase Center and the Emergence of the Translation System. *Life* **2017**, *7*, 21. [[CrossRef](#)]
22. Demongeot, J.; Seligmann, H. Evolution of tRNA into rRNA secondary structures. *Gene Rep.* **2019**, *17*, 100483. [[CrossRef](#)]
23. Sayers, E.W.;avanaugh, M.; Clark, K.; Ostell, J.; Pruitt, K.D.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Res.* **2019**, *48*, D84–D86.
24. Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443–453. [[CrossRef](#)]
25. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [[CrossRef](#)]
26. Thompson, J.D.; Gibson, T.J.; Higgins, D.G. Multiple Sequence Alignment Using ClustalW and ClustalX. *Curr. Protoc. Bioinform.* **2002**, *00*, 2.3.1–2.3.22. [[CrossRef](#)] [[PubMed](#)]
27. Spang, A.; Stairs, C.W.; Dombrowski, N.; Eme, L.; Lombard, J.; Caceres, E.F.; Greening, C.; Baker, B.J.; Ettema, T.J.G. Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* **2019**, *4*, 1138–1148. [[CrossRef](#)] [[PubMed](#)]
28. Spang, A.; Saw, J.H.; Jørgensen, S.L.; Zaremba-Niedzwiedzka, K.; Martijn, J.; Lind, A.E.; Van Eijk, R.; Schleper, C.; Guy, L.; Ettema, T.J.G. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **2015**, *521*, 173–179. [[CrossRef](#)]
29. Crooks, G.E.; Hon, G.; Chandonia, J.-M.; Brenner, S.E. WebLogo: A Sequence Logo Generator. *Genome Res.* **2004**, *14*, 1188–1190. [[CrossRef](#)]
30. Zamudio, G.S.; José, M.V. Identity Elements of tRNA as Derived from Information Analysis. *Orig. Life Evol. Biosph.* **2017**, *48*, 73–81. [[CrossRef](#)]
31. Meilã, M. Comparing Clusterings by the Variation of Information. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, 24–27 August 2003: Proceedings*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 173–187, ISBN 978-3-540-40720-1.



32. Zamudio, G.S.; Palacios-Pérez, M.; José, M.V. Information theory unveils the evolution of tRNA identity elements in the three domains of life. *Theory Biosci.* **2020**, *139*, 77–85. [[CrossRef](#)]
33. Bernier, C.R.; Petrov, A.S.; Waterbury, C.C.; Jett, J.; Li, F.; Freil, L.E.; Xiong, X.; Wang, L.; Migliozi, B.; Hershkovits, E.; et al. RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss.* **2014**, *169*, 195–207. [[CrossRef](#)]
34. Reuter, J.S.; Mathews, D.H. RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinform.* **2010**, *11*, 129. [[CrossRef](#)] [[PubMed](#)]
35. Kerpedjiev, P.; Hammer, S.; Hofacker, I.L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **2015**, *31*, 3377–3379. [[CrossRef](#)] [[PubMed](#)]
36. Okonechnikov, K.; Golosova, O.; Fursov, M. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **2012**, *28*, 1166–1167. [[CrossRef](#)] [[PubMed](#)]
37. Rother, M.; Rother, K.; Puton, T.; Bujnicki, J.M. ModeRNA: A tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.* **2011**, *39*, 4007–4022. [[CrossRef](#)]
38. Gong, S.; Zhang, C.; Zhang, Y. RNA-align: Quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics* **2019**, *35*, 4459–4461. [[CrossRef](#)]
39. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem* **2004**, *25*, 1605–1612. [[CrossRef](#)]
40. Prosdocimi, F.; Linard, B.; Pontarotti, P.; Poch, O.; Thompson, J.D. Controversies in modern evolutionary biology: The imperative for error detection and quality control. *BMC Genom.* **2012**, *13*, 5. [[CrossRef](#)]
41. Schroeder, R.; Barta, A.; Semrad, K. Strategies for RNA folding and assembly. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 908–919. [[CrossRef](#)]
42. Schultes, E.A.; Bartel, D.P. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* **2000**, *289*, 448–452. [[CrossRef](#)]
43. Brion, P.; Westhof, E. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 113–137. [[CrossRef](#)]
44. Draper, D.E. A guide to ions and RNA structure. *RNA* **2004**, *10*, 335–343. [[CrossRef](#)] [[PubMed](#)]
45. De Farias, S.T. Suggested phylogeny of tRNAs based on the construction of ancestral sequences. *J. Theor. Biol.* **2013**, *335*, 245–248. [[CrossRef](#)] [[PubMed](#)]
46. Jühling, F.; Mörl, M.; Hartmann, R.K.; Sprinzl, M.; Stadler, P.F.; Pütz, J. tRNADB 2009: Compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **2009**, *37*, D159–D162. [[CrossRef](#)] [[PubMed](#)]
47. Korostelev, A.; Trakhanov, S.; Laurberg, M.; Noller, H.F. Crystal Structure of a 70S Ribosome-tRNA Complex Reveals Functional Interactions and Rearrangements. *Cell* **2006**, *126*, 1065–1077. [[CrossRef](#)]
48. Krupkin, M.; Matzov, D.; Tang, H.; Metz, M.; Kalaora, R.; Belousoff, M.J.; Zimmerman, E.; Bashan, A.; Yonath, A. A vestige of a prebiotic bonding machine is functioning within the contemporary ribosome. *Philos. Trans. R. Soc. B Biol. Sci.* **2011**, *366*, 2972–2978. [[CrossRef](#)]
49. De Farias, S.T.; José, M.V. Transfer RNA: The molecular demiurge in the origin of biological systems. *Prog. Biophys. Mol. Boil.* **2020**, *153*, 28–34. [[CrossRef](#)]
50. Prosdocimi, F.; de Farias, S.T. From FUCA To LUCA: A Theoretical Analysis on the Common Descent of Gene Families. *Acta Sci. Microbiol.* **2020**, *3*, 1–9.
51. Ivanov, V.I.; Bondarenko, S.A.; Zdobnov, E.M.; Beniaminov, A.D.; Minyat, E.E.; Ulyanov, N.B. A pseudoknot-compatible universal site is located in the large ribosomal RNA in the peptidyltransferase center. *FEBS Lett.* **1999**, *446*, 60–64. [[CrossRef](#)]
52. Polacek, N.; Mankin, A.S. The Ribosomal Peptidyl Transferase Center: Structure, Function, Evolution, Inhibition. *Crit. Rev. Biochem. Mol. Biol.* **2005**, *40*, 285–311. [[CrossRef](#)]
53. Chirkova, A.; Erlacher, M.D.; Clementi, N.; Żywicki, M.; Aigner, M.; Polacek, N. The role of the universally conserved A2450–C2063 base pair in the ribosomal peptidyl transferase center. *Nucleic Acids Res.* **2010**, *38*, 4844–4855. [[CrossRef](#)]
54. Davidovich, C.; Belousoff, M.J.; Wekselman, I.; Shapira, T.; Krupkin, M.; Zimmerman, E.; Bashan, A.; Yonath, A. The Proto-Ribosome: An ancient nano-machine for peptide bond formation. *Isr. J. Chem.* **2010**, *50*, 29–35. [[CrossRef](#)] [[PubMed](#)]
55. Terasaka, N.; Hayashi, G.; Katoh, T.; Suga, H. An orthogonal ribosome-tRNA pair via engineering of the peptidyl transferase center. *Nat. Chem. Biol.* **2014**, *10*, 555–557. [[CrossRef](#)] [[PubMed](#)]

56. Lehmann, J. Physico-chemical Constraints Connected with the Coding Properties of the Genetic System. *J. Theor. Biol.* **2000**, *202*, 129–144. [[CrossRef](#)] [[PubMed](#)]
57. Lehmann, J.; Cibils, M.; Libchaber, A. Emergence of a Code in the Polymerization of Amino Acids along RNA Templates. *PLoS ONE* **2009**, *4*, e5773. [[CrossRef](#)] [[PubMed](#)]
58. Zaucha, J.; Heddle, J.G. Resurrecting the Dead (Molecules). *Comput. Struct. Biotechnol. J.* **2017**, *15*, 351–358. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

## 7. SÍNTESIS DE *EXTRA*

La importancia de los tRNA radica en su función como adaptadores en el proceso de traducción, transfiriendo las instrucciones contenidas en el código genético, mediante el anticodón, a la incorporación de los aminoácidos para la adecuada formación de proteínas. Cada tRNA es activado por una aaRL específica que, sin leer el anticodón, carga el aa correcto correspondiente; el discernimiento se basa en el reconocimiento positivo y negativo de ciertos nucleótidos específicos, llamados elementos de identidad, que se encuentran a lo largo del tRNA. Utilizando análisis de la teoría de la información se examinaron las secuencias de tRNA existentes, para saber si hay alguna covariación entre la información contenida en el anticodón y otras posiciones nucleotídicas que pudieran constituir precisamente elementos de identidad (Zamudio et al., 2020).

Las secuencias de tRNA, depositadas en las bases de datos disponibles más actualizadas y curadas, se clasificaron inicialmente según los tres grandes dominios: bacteria, arquea y eucaria; además los isoaceptores se agruparon según los 20 aa canónicos con que son activados. Los isoaceptores que no poseen la longitud más común fueron descartados, para obtener tRNAs con el mismo número de nt, lo que los hizo comparables y se representaron entonces como una única secuencia con 76 sitios. Terminando con un total de 60 'secuencias' (Zamudio et al., 2020).

Al analizar por grupos los nt de las secuencias de tRNA, se encontró correlación de información entre el anticodón y otros sitios, que por ende constituirían elementos de identidad; se encontraron además otros distintos grupos de nt con relación entre sí, pero sin relación informativa con el anticodón. Los tRNA fueron ordenados de acuerdo con un diagrama ramificado (dendrograma), que agrupa los isoaceptores por el número de sitios informativos con relación al anticodón; interesantemente no se observó relación alguna entre los tRNA agrupados por cantidad de información y las características fisicoquímicas o estructurales de los aa o las aaRL correspondientes (Zamudio *et al.*, 2020, parte en que colaboré mayormente con la descripción de dichas características de aa y aaRLs, para observar si ello tenía implicaciones en la posición o número de elementos de identidad de los isoaceptores).

Los grupos de identidad encontrados para todos los isoaceptores se compararon intra- e interdominio para determinar un patrón de diferenciación y así arrojar luz sobre su evolución. La diferencia más notable radicó en la cantidad de sitios informativos por dominio biológico, cuyo número incrementa iniciando con bacteria, luego eucaria y finalmente arquea, lo que no es acorde con la filogenia comúnmente aceptada en que los eucariontes son los organismos más complejos o al menos más recientes. Porque aunque las bacterias sí tienen el menor número de sitios informativos con respecto al anticodón, las arqueas poseen la mayor cantidad de elementos de identidad; lo que probablemente se deba a que las arqueas, en varios casos y a diferencia de los eucariontes, carecen de brazos T y D, de manera que el alto número de posiciones podría servir para asegurar la probabilidad de aún preservar sitios de identidad que aseguren el cargado del aminoácido correcto (Zamudio et al., 2020).

Además de los tRNA, el PTC es el RNA más importante del proceso de traducción. Con el objetivo de dilucidar la importancia de la posición de cada uno de los nucleótidos, se utilizaron análisis de la teoría de la información para analizar la secuencia de los PTC, además se observaron los sitios completamente conservados para todos los PTC (PTC-all), para los de bacterias (PTC-bac) y los de arqueas (PTC-arc) y se reconstruyeron las estructuras correspondientes, utilizando el PTC de *Thermus thermophilus* (PTC-Ther2) como referencia (Prosdocimi et al., 2020).

Primeramente se descargaron las secuencias completas del 23SrRNA de bacterias y arqueas, que se alinearon para obtener el *peptidyl transferase centre*, utilizando la secuencia de PTC-Ther2 (tomada del 23SrRNA de Ther2, PDB id 4V4I) como cebador; las secuencias obtenidas se realinearon entre sí para obtener un alineamiento múltiple de todas las secuencias (*multiple sequence alignment*, MSA) con la menor cantidad de gaps; se obtuvo entonces un 'conjunto completo' de todos los PTC, dicho conjunto se dividió acorde con los dos grandes dominios, los PTC de bacterias se subdividieron además según las familias con mayor número de representatividad, mientras que los PTC de arqueas se dejaron sin subclasificar al no encontrar suficientes representantes diferentes (Prosdocimi et al., 2020).

Para cada una de las 6 divisiones se observó el número de los sitios 100% conservados y el número y tamaño de gaps, entre otras características particulares de cada conjunto (Prosdocimi et al., 2020). Los MSA de PTC-all, PTC-bac y PTC-arc se representaron en formato logo, lo que permitió visualizar gráficamente las posiciones de los nt conservados y característicos de cada conjunto, revelando los abundantes sitios conservados en todos los PTC y los nt específicos de cada conjunto; mientras que el sitio A2451, esencial para la formación del enlace peptídico, es totalmente invariante (en que colaboré). Se obtuvieron secuencias consenso de cada uno de los conjuntos mencionados (PTC-all, PTC-bac y PTC-arc) y se reconstruyó su estructura terciaria tomando como plantilla la estructura experimental del PTC-Ther2; las posiciones 100% conservadas en cada conjunto fueron mapeadas en las estructuras reconstruidas (que yo elaboré).

El PTC-Ther2 se utilizó como referente para los ulteriores análisis; por un lado se tomó su estructura secundaria y terciaria extrayéndola de la representación completa del 23SrRNA determinada experimentalmente, por otro lado se reconstruyó *de novo* la estructura secundaria utilizando programas de predicción (procedimientos en que colaboré); sin embargo es probable que ambas conformaciones existiesen prebióticamente, o como parte del Primer Ancestro Común Universal (*First Universal Common Ancestor*, FUCA), como respuesta a la presencia de diversos ligandos y otras condiciones ambientales, dado que se requiere la ruptura de únicamente tres pares Watson-Crick para tener una u otra (Prosdocimi et al., 2020), como analogía de isómeros químicos.

Los MSA de los PTC se utilizaron para llevar a cabo análisis de teoría de la información, para revelar la relación mutua de cada sitio nucleotídico con respecto a los demás con base en la variación de la información. Dichos análisis revelaron grupos bien definidos de nucleótidos perfectamente covariantes; de tal manera que el conjunto PTC-all reveló 25 posiciones de nucleótidos, el conjunto PTC-bac presentó 71 posiciones y el conjunto PTC-arc presentó 101 posiciones. Los grupos de nucleótidos covariantes para cada conjunto (PTC-all, PTC-bac y PTC-arc) se mapearon en el PTC-Ther2, tanto en la estructura secundaria, experimental y reconstruida, como en la estructura terciaria experimental, donde se aprecia el sitio A2451 o A12 del modelo actual (en que colaboré con el mapeo). De manera notable, utilizando la estructura secundaria del

PTC-Ther2, es posible visualizar que los distintos grupos de nucleótidos covariantes en el conjunto PTC-Bac, se corresponden con porciones de 5 proto-tRNAs bien definidos, de manera que los sitios informativos fungieron como una guía para la formación del PTC moderno (Prosdocimi et al., 2020).

Finalmente, se comparó la estructura tridimensional experimental del PTC-Ther2 y las reconstrucciones de las secuencias consenso de PTC-all, PTC-bac y PTC-arc, lo que permitió observar las sutiles, pero importantes diferencias que caracterizan cada tipo de PTC (Prosdocimi et al. 2020).

El origen y la evolución de la maquinaria de traducción es una de las principales transiciones evolutivas y dicha maquinaria no puede entenderse sin la evolución misma del ARN (RNA), que pudo haber configurado el código genético primitivo, que seguía un patrón RNY (Eigen, 1971; Eigen and Schuster, 1978), que evolucionó mediante la ruta dicotómica compuesta por Ex1 y Ex2, que se complementaron y finalmente formaron el SGC actual (José et al., 2009; José et al., 2011).

Con base en la ruta descrita, de manera análoga a la reconstrucción del proteoma desde los CSBSs hasta los proteomas extendidos, fue posible rastrear la evolución del **RNAoma**. Descubrimos que algunas porciones de los tres tipos de ARN ribosómico (ARNr) y porciones de ciertas moléculas de tRNA fueron las primeras en codificarse y probablemente se plegaban como horquillas. Finalmente, los rRNA y tRNA completos, así como el componente de RNA de la ribonucleasa P, el 6S RNA y el dual tmRNA menos conocidos, fueron codificados por tripletes pertenecientes a ambos ExGCs; aunque las porciones codificadas por tripletes de cada ExGC son complementarias entre sí y pueden haberse ensamblado de manera cooperativa hasta finalmente constituir las moléculas modernas de RNA (Palacios-Pérez & José, en preparación).

## 8. CONCLUSIONES Y PROSPECTIVAS

Como hemos podido observar, es posible encontrar, a manera de fotografías de momentos evolutivos congelados, vestigios de componentes prebióticos y proto-bióticos en las formas de vida actuales; para lo cual se han utilizado enfoques diversos que van desde los análisis más puramente matemáticos, como grupos de renormalización, teoría de grafos o teoría de la información, hasta combinaciones de varias técnicas bioinformáticas analizando genomas completos.

Una manera de analizar dichos vestigios con base en un código genético; así tenemos el código genético primitivo (PGC) constituido por cadenas ribonucleotídicas del tipo RNY (Eigen, 1971; Eigen and Schuster, 1978), el cual codifica para 8 aa que a su vez pueden ordenarse según su requerimiento polar, revelando así el punto en el cual el ordenamiento se quiebra para dar origen al siguiente paso evolutivo (José et al., 2015). Los aa codificados por el PGC forman a su vez una colección de péptidos capaces de unir las diversas moléculas formadas prebióticamente en la Tierra primitiva, los CSBSs (Palacios-Pérez et al., 2018).

El siguiente paso evolutivo que siguió el PGC es una ruta dicotómica que, mediante grupos de renormalización, se demostró algebraicamente y es biológicamente plausible; de esta manera emergieron dos códigos genéticos extendidos (ExGCs) a partir de dos tipos de mutaciones del patrón RNY. Dichos ExGCs se complementan mutuamente para dar origen al código genético estándar (SGC) del LUCA (José et al., 2009; José et al., 2011); de la misma manera se observa que los fragmentos codificados por uno u otro tipo de ExGC pertenecen a porciones distintas y complementarias de las proteínas modernas, que sin aún formarse todavía, podían probablemente realizar, aunque quizá deficientemente, las funciones proteicas correspondientes y por ende conferir un esbozo de lo que más adelante sería el proteoma completo del LUCA (Palacios-Pérez and José, 2019).

Con base en la ruta evolutiva propuesta partiendo de cadenas de RNA con un patrón RNY y posteriormente el surgimiento de los códigos extendidos, es posible también rastrear la evolución del RNA de manera análoga a la formación de proteínas (Palacios-Pérez & José, en preparación).

Después de la propuesta de un PGC basado en cadenas indeterminadas de ribonucleótidos, se planteó que las moléculas hereditarias iniciales fueron específicamente proto-tRNAs, pues son de tamaño compatible con las condiciones de polimerización prebiótica y pudieron al mismo tiempo fungir como mensajeros en el primitivo proceso de traducción (Eigen and Winkler-Oswatitsch, 1981a; Eigen and Winkler-Oswatitsch, 1981b); a su vez, los proto-tRNA formaron el PTC, sitio catalítico del ribosoma fundamental *sine qua non* para la traducción (Farias et al., 2014; De Farias et al., 2016). Dicho PTC es formado efectivamente por proto-tRNAs, que se ensamblan según los grupos de información contenida en las mismas secuencias; visto de otra manera, en la variación de la información proveída por los nt del PTC, se observan los vestigios del ensamblaje de los proto-tRNA que le dieron forma. Adicionalmente, es posible documentar la conformación del PTC del LUCA, así como observar los sutiles cambios –tanto a nivel de secuencia, de estructura secundaria y de estructura terciaria–, que darán origen al PTC de arqueas y al PTC de bacterias (Prosdocimi et al., 2020)

Otra de las moléculas indispensables en la traducción moderna de proteínas son los tRNA, que deben ser activados con el correcto aa, codificado por el anticodón; sin embargo, dicho aa es cargado por proteínas (aaRL) que reconocen sitios específicos (elementos de identidad), lejos del anticodón, a lo largo de la secuencia de los tRNA. Con base en la teoría de la información, se pueden averiguar los elementos de identidad que se correlacionan informacionalmente con el anticodón y serían los sitios reconocidos por las aaRL, sin importar las características fisicoquímicas de los aminoácidos (Zamudio et al., 2020).

Como podemos observar, hay varias maneras en que podemos encontrar vestigios ancestrales en organismos modernos e intentar delinear las cualidades esenciales que definieron la vida en sus comienzos y la comprensión de cómo evolucionaron esas características desde el FUCA (Prosdocimi and Farias, 2020; Prosdocimi et al., 2019) hacia la formación del repertorio completo del LUCA e incluso esbozar las primeras particularidades de organismos más diferenciados.

Finalmente, este tipo de trabajos podrían guiar la reconstrucción experimental de proteínas antiguas o sus predecesores y explicar la evolución de la biogeoquímica (Carter, 2014; Hochberg and Thornton, 2017; Garcia and Kaçar, 2019).



## 9. REFERENCIAS BIBLIOHEMEROGRÁFICAS

- Agmon, I. C.** (2016). Could a Proto-Ribosome Emerge Spontaneously in the Prebiotic World? *Molecules* **21**, 1701.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J.** (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.
- Bray, M. S., Lenz, T. K., Haynes, J. W., Bowman, J. C., Petrov, A. S., Reddi, A. R., Hud, N. V., Williams, L. D. and Glass, J. B.** (2018). Multiple prebiotic metals mediate translation. *PNAS* **115**, 12164–12169.
- Brisson, D., Vohl, M.-C., St-Pierre, J., Hudson, T. J. and Gaudet, D.** (2001). Glycerol: a neglected variable in metabolic processes? *BioEssays* **23**, 534–542.
- Campbell, T. D., Febrian, R., McCarthy, J. T., Kleinschmidt, H. E., Forsythe, J. G. and Bracher, P. J.** (2019). Prebiotic condensation through wet-dry cycling regulated by deliquescence. *Nat Commun* **10**, 1–7.
- Carter, C. W.** (2014). Urzymology: experimental access to a key transition in the appearance of enzymes. *J. Biol. Chem.* **289**, 30213–30220.
- Cech, T. R.** (2009). Evolution of biological catalysis: ribozyme to RNP enzyme. *Cold Spring Harb. Symp. Quant. Biol.* **74**, 11–16.
- Cech, T. R.** (2012). The RNA worlds in context. *Cold Spring Harb Perspect Biol* **4**, a006742.
- Chatterjee, S.** (2016). A symbiotic view of the origin of life at hydrothermal impact crater-lakes. *Phys. Chem. Chem. Phys.* **18**, 20033–20046.
- Cornell, C. E., Black, R. A., Xue, M., Litz, H. E., Ramsay, A., Gordon, M., Mileant, A., Cohen, Z. R., Williams, J. A., Lee, K. K., et al.** (2019). Prebiotic amino acids bind to and stabilize prebiotic fatty acid membranes. *PNAS* **116**, 17239–17244.
- Cornish-Bowden, A. and Cárdenas, M. L.** (2017). Life before LUCA. *Journal of Theoretical Biology* **434**, 68–74.
- Cronin, J. R. and Moore, C. B.** (1971). Amino Acid Analyses of the Murchison, Murray, and Allende Carbonaceous Chondrites. *Science* **172**, 1327–1329.
- Damer, B. and Deamer, D.** (2019). The Hot Spring Hypothesis for an Origin of Life. *Astrobiology* **20**, 429–452.
- Darwin, C.** (1859). *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life* /. London: John Murray,.
- De Farias, S. T., Rêgo, T. G. and José, M. V.** (2016). tRNA Core Hypothesis for the Transition from the RNA World to the Ribonucleoprotein World. *Life* **6**, 15.
- Delaye, L., Becerra, A. and Lazcano, A.** (2005). The Last Common Ancestor: What's in a name? *Orig Life Evol Biosph* **35**, 537–554.
- Di Giulio, M.** (1997). On the RNA world: evidence in favor of an early ribonucleopeptide world. *J. Mol. Evol.* **45**, 571–578.
- Di Giulio, M.** (2011). The Last Universal Common Ancestor (LUCA) and the Ancestors of Archaea and Bacteria were Progenotes. *J Mol Evol* **72**, 119–126.
- Eigen, M.** (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523.
- Eigen, M. and Schuster, P.** (1978). The Hypercycle - A principle of natural self-organization Part C: The realistic hypercycle. *Naturwissenschaften* **65**, 341–369.
- Eigen, M. and Winkler-Oswatitsch, R.** (1981a). Transfer-RNA: The early adaptor. *Naturwissenschaften* **68**, 217–228.
- Eigen, M. and Winkler-Oswatitsch, R.** (1981b). Transfer-RNA, an early gene? *Naturwissenschaften* **68**, 282–292.
- Farias, S. T., Rêgo, T. G. and José, M. V.** (2014). Origin and evolution of the Peptidyl Transferase Center from proto-tRNAs. *FEBS Open Bio* **4**, 175–178.
- Fedor, M. J.** (2002). The role of metal ions in RNA catalysis. *Curr. Opin. Struct. Biol.* **12**, 289–295.
- Fenotipo | NHGRI** [Genome.gov](http://Genome.gov).
- Forterre, P. and Gribaldo, S.** (2007). The origin of modern terrestrial life. *HFSP Journal* **1**, 156–168.
- Forterre, P., Gribaldo, S. and Brochier, C.** (2005). [Luca: the last universal common ancestor]. *Med Sci (Paris)* **21**, 860–865.
- Garcia, A. K. and Kaçar, B.** (2019). How to resurrect ancestral proteins as proxies for ancient biogeochemistry. *Free Radical Biology and Medicine* **140**, 260–269.
- Gesteland, R. F., Cech, T. and Atkins, J. F.** (2006). *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*. Cold Spring Harbor Laboratory Press.
- Gil, R., Silva, F. J., Peretó, J. and Moya, A.** (2004). Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.* **68**, 518–537, table of contents.
- Gilbert, W.** (1986). Origin of life: The RNA world. *Nature* **319**, 618.
- Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Hutchison, C. A., Smith, H. O. and Venter, J. C.** (2006). Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 425–430.
- Goldman, A. D., Bernhard, T. M., Dolzhenko, E. and Landweber, L. F.** (2013). LUCApedia: a database for the study of ancient life. *Nucleic Acids Res* **41**, D1079–D1082.
- Goto, N., Kurokawa, K. and Yasunaga, T.** (2007). Analysis of invariant sequences in 266 complete genomes. *Gene* **401**, 172–180.
- Griffin, L., West, D. J. and West, B. J.** (2000). Random stride intervals with memory. *J Biol Phys* **26**, 185–202.
- Hochberg, G. K. A. and Thornton, J. W.** (2017). Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annual Review of Biophysics* **46**, 247–269.
- Huang, F., Bugg, C. W. and Yarus, M.** (2000). RNA-Catalyzed CoA, NAD, and FAD synthesis from phosphopantetheine, NMN, and FMN. *Biochemistry* **39**, 15548–15555.

- Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., et al.** (2020). Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature* **577**, 519–525.
- Itaya, M.** (1995). An estimation of minimal genome size required for life. *FEBS Lett.* **362**, 257–260.
- Jadhav, V. R. and Yarus, M.** (2002). Coenzymes as coribozymes. *Biochimie* **84**, 877–888.
- Janas, T., Janas, T. and Yarus, M.** (2006). Specific RNA binding to ordered phospholipid bilayers. *Nucleic Acids Res.* **34**, 2128–2136.
- Jensen, R. A.** (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425.
- Ji, H.-F., Kong, D.-X., Shen, L., Chen, L.-L., Ma, B.-G. and Zhang, H.-Y.** (2007). Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biology* **8**, R176.
- José, M. V., Govezensky, T., García, J. A. and Bobadilla, J. R.** (2009). On the Evolution of the Standard Genetic Code: Vestiges of Critical Scale Invariance from the RNA World in Current Prokaryote Genomes. *PLOS ONE* **4**, e4340.
- José, M. V., Morgado, E. R. and Govezensky, T.** (2011). Genetic Hotels for the Standard Genetic Code: Evolutionary Analysis Based upon Novel Three-Dimensional Algebraic Models. *Bull Math Biol* **73**, 1443–1476.
- José, M. V., Zamudio, G. S., Palacios-Pérez, M., Bobadilla, J. R. and de Farias, S. T.** (2015). Symmetrical and Thermodynamic Properties of Phenotypic Graphs of Amino Acids Encoded by the Primeval RNY Code. *Orig Life Evol Biosph* **45**, 77–83.
- Koonin, E. V.** (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136.
- Koonin, E. V. and Novozhilov, A. S.** (2009). Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* **61**, 99–111.
- Lancet, D., Zidovetzki, R. and Markovitch, O.** (2018). Systems protobiology: origin of life in lipid catalytic networks. *J R Soc Interface* **15**, .
- Lazcano, A. and Bada, J. L.** (2003). The 1953 Stanley L. Miller Experiment: Fifty Years of Prebiotic Organic Chemistry. *Orig Life Evol Biosph* **33**, 235–242.
- Lazcano, A. and Miller, S. L.** (1999). On the origin of metabolic pathways. *J. Mol. Evol.* **49**, 424–431.
- Lupas, A. N. and Alva, V.** (2017). Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *Journal of Structural Biology* **198**, 74–81.
- Ma, B.-G., Chen, L., Ji, H.-F., Chen, Z.-H., Yang, F.-R., Wang, L., Qu, G., Jiang, Y.-Y., Ji, C. and Zhang, H.-Y.** (2008). Characters of very ancient proteins. *Biochemical and Biophysical Research Communications* **366**, 607–611.
- Mariscal, C., Barahona, A., Aubert-Kato, N., Aydinoglu, A. U., Bartlett, S., Cárdenas, M. L., Chandru, K., Cleland, C., Cocanougher, B. T., Comfort, N., et al.** (2019). Hidden Concepts in the History and Philosophy of Origins-of-Life Studies: a Workshop Report. *Orig Life Evol Biosph* **49**, 111–145.
- McInerney, J. O.** (2016). Evolution: A four billion year old metabolism. *Nature Microbiology* **1**, 16139.
- Medini, D., Donati, C., Tettelin, H., Masignani, V. and Rappuoli, R.** (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594.
- Miller, S. L.** (1953). A Production of Amino Acids Under Possible Primitive Earth Conditions. *Science* **117**, 528–529.
- Miller, S. L., Urey, H. C. and Oró, J.** (1976). Origin of organic compounds on the primitive earth and in meteorites. *J. Mol. Evol.* **9**, 59–72.
- Mushegian, A. R. and Koonin, E. V.** (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10268–10273.
- NCBI Taxonomy Homepage.**
- Pace, N. R., Sapp, J. and Goldenfeld, N.** (2012). Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proc Natl Acad Sci U S A* **109**, 1011–1018.
- Palacios-Pérez, M. and José, M. V.** (2019). The evolution of proteome: From the primeval to the very dawn of LUCA. *BioSystems* **181**, 1–10.
- Palacios-Pérez, M., Andrade-Díaz, F. and José, M. V.** (2018). A Proposal of the Ur-proteome. *Orig Life Evol Biosph* **48**, 245–258.
- Parker, E. T., Cleaves, J. H., Burton, A. S., Glavin, D. P., Dworkin, J. P., Zhou, M., Bada, J. L. and Fernández, F. M.** (2014). Conducting miller-urey experiments. *J Vis Exp* e51039.
- Preiner, M., Asche, S., Becker, S., Betts, H. C., Boniface, A., Camprubi, E., Chandru, K., Erastova, V., Garg, S. G., Khawaja, N., et al.** (2020). The Future of Origin of Life Research: Bridging Decades-Old Divisions. *Life* **10**, 20.
- Prosdocimi, F. and Farias, S. T. de** (2020). From FUCA To LUCA: A Theoretical Analysis on the Common Descent of Gene Families. *Acta Scientific Microbiology* **3**, .
- Prosdocimi, F., José, M. V. and de Farias, S. T.** (2019). The First Universal Common Ancestor (FUCA) as the Earliest Ancestor of LUCA's (Last UCA) Lineage. In *Evolution, Origin of Life, Concepts and Methods* (ed. Pontarotti, P.), pp. 43–54. Cham: Springer International Publishing.
- Prosdocimi, F., Zamudio, G. S., Palacios-Pérez, M., Torres de Farias, S. and José, M. V.** (2020). The Ancient History of Peptidyl Transferase Center Formation as Told by Conservation and Information Analyses. *Life* **10**, 134.
- Raanan, H., Poudel, S., Pike, D. H., Nanda, V. and Falkowski, P. G.** (2020). Small protein folds at the root of an ancient metabolic network. *PNAS* **117**, 7193–7199.
- Ranea, J. A. G., Sillero, A., Thornton, J. M. and Orengo, C. A.** (2006). Protein Superfamily Evolution and the Last Universal Common Ancestor (LUCA). *J Mol Evol* **63**, 513–525.
- Raymann, K., Brochier-Armanet, C. and Gribaldo, S.** (2015). The two-domain tree of life is linked to a new root for the Archaea. *PNAS* **112**, 6670–6675.
- Ribeiro, A. J. M., Tyzack, J. D., Borkakoti, N., Holliday, G. L. and Thornton, J. M.** (2020). A global analysis of function and conservation of catalytic residues in enzymes. *J. Biol. Chem.* **295**, 314–324.
- Risso, V. A., Martínez-Rodríguez, S., Candel, A. M., Krüger, D. M., Pantoja-Uceda, D., Ortega-Muñoz, M., Santoyo-Gonzalez, F., Gaucher, E. A., Kamerlin, S. C. L., Bruix, M., et al.** (2017). De novo active sites for resurrected Precambrian enzymes. *Nature Communications* **8**, 16113.

- Rodin, A. S., Rodin, S. N. and Carter, C. W.** (2009). On primordial sense-antisense coding. *J Mol Evol* **69**, 555–567.
- Schroeder, M.** (1991). *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. First ed. N.Y., U.S.A: W H Freeman & Co (Sd).
- Segré, D. and Lancet, D.** (2000). Composing life. *EMBO Rep.* **1**, 217–222.
- Segré, D., Ben-Eli, D., Deamer, D. W. and Lancet, D.** (2001). The lipid world. *Orig Life Evol Biosph* **31**, 119–145.
- Sharov, A. A.** (2009). Coenzyme autocatalytic network on the surface of oil microspheres as a model for the origin of life. *Int J Mol Sci* **10**, 1838–1852.
- Sharov, A. A.** (2016). Coenzyme world model of the origin of life. *Biosystems* **144**, 8–17.
- Sobolevsky, Y. and Trifonov, E. N.** (2006). Protein modules conserved since LUCA. *J. Mol. Evol.* **63**, 622–634.
- Sobolevsky, Y., Guimarães, R. C. and Trifonov, E. N.** (2013). Towards functional repertoire of the earliest proteins. *J. Biomol. Struct. Dyn.* **31**, 1293–1300.
- Spang, A. and Ettema, T. J. G.** (2016). Microbial diversity: The tree of life comes of age. *Nature Microbiology* **1**, 1–2.
- Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L. and Ettema, T. J. G.** (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179.
- Szathmáry, E.** (1990). Towards the evolution of ribozymes. *Nature* **344**, 115.
- Szathmáry, E.** (1993). Coding coenzyme handles: a hypothesis for the origin of the genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 9916–9920.
- Szathmáry, E., Szathmáry, E. and Szathmáry, E.** (1999). The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends in Genetics* **15**, 223–229.
- Tamura, K. and Schimmel, P.** (2003). Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptors. *PNAS* **100**, 8666–8669.
- Tettelin, H., Massignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., et al.** (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13950–13955.
- Trifonov, E. N.** (2004). The triplet code from first principles. *J. Biomol. Struct. Dyn.* **22**, 1–11.
- Trifonov, E. N.** (2006). Self-inflicted Fear of Evolution. *Orig Life Evol Biosph* **36**, 557–558.
- Urey, H. C.** (1966). A review of evidence for biological material in meteorites. *Life Sci Space Res* **4**, 35–59.
- Wächtershäuser, G.** (1988a). Before enzymes and templates: theory of surface metabolism. *Microbiol. Rev.* **52**, 452–484.
- Wächtershäuser, G.** (1988b). Before enzymes and templates: theory of surface metabolism. *Microbiology and Molecular Biology Reviews* **52**, 452–484.
- Wächtershäuser, G.** (1990). Evolution of the first metabolic cycles. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 200–204.
- Wächtershäuser, G.** (2014). The Place of RNA in the Origin and Early Evolution of the Genetic Machinery. *Life* **4**, 1050–1091.
- Watanabe, K. and Suzuki, T.** (2001). Genetic Code and its Variants. In *eLS*, p. American Cancer Society.
- Watanabe, K. and Suzuki, T.** (2008). Universal Genetic Code and its Natural Variations. In *eLS*, p. American Cancer Society.
- Weber, A. L.** (1989). Model of early self-replication based on covalent complementarity for a copolymer of glycerate-3-phosphate and glycerol-3-phosphate. *Orig Life Evol Biosph* **19**, 179–186.
- Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S. and Martin, W. F.** (2016). The physiology and habitat of the last universal common ancestor. *Nature Microbiology* **1**, 16116.
- Williams, T. A., Foster, P. G., Nye, T. M. W., Cox, C. J. and Embley, T. M.** (2012). A congruent phylogenomic signal places eukaryotes within the Archaea. *Proceedings of the Royal Society B: Biological Sciences* **279**, 4870–4879.
- Wilson, K. G.** (1979). Problems in Physics with many Scales of Length. *Sci Am* **241**, 158–179.
- Woese, C. R.** (1964). Universality in the genetic code. *Science* **144**, 1030–1031.
- Woese, C.** (1998). The universal ancestor. *Proc Natl Acad Sci U S A* **95**, 6854–6859.
- Woese, C. R.** (2000). Interpreting the universal phylogenetic tree. *PNAS* **97**, 8392–8396.
- Woese, C. R. and Fox, G. E.** (1977). The concept of cellular evolution. *J. Mol. Evol.* **10**, 1–6.
- Yarus, M.** (2002). Primordial genetics: phenotype of the ribocyte. *Annu. Rev. Genet.* **36**, 125–151.
- Yarus, M.** (2011). Getting past the RNA world: the initial Darwinian ancestor. *Cold Spring Harb Perspect Biol* **3**,.
- Yčas, M.** (1974). On earlier states of the biochemical system. *Journal of Theoretical Biology* **44**, 145–160.
- Zamudio, G. S., Palacios-Pérez, M. and José, M. V.** (2020). Information theory unveils the evolution of tRNA identity elements in the three domains of life. *Theory Biosci.* **139**, 77–85.



## **ESTO PASARÁ**

Cantante Eugenia León  
Letra Gutiérrez Mueller

El miedo de hoy mañana será templanza,  
el llanto será esperanza  
y evolución

Estamos en pausa, pero el amor avanza  
En tiempos desoladores,  
por dentro nos crecen flores  
y compasión

Esto pasará, esto pasará  
Tenemos el cielo intacto y la tormenta terminará  
Esto pasará, esto pasará  
Es sólo cuestión de tiempo, la música del silencio renacerá  
Esto pasará, esto pasará

Lo que hoy es incertidumbre será sonrisa  
Y se acabará este ciclo de obscuridad  
Mañana se irán los muros, vendrá la brisa  
Y ya nunca olvidaremos, que cada segundo es una oportunidad

Esto pasará, esto pasará  
Tenemos el cielo intacto y la tormenta terminará  
Esto pasará, esto pasará  
Es sólo cuestión de tiempo, la música del silencio renacerá

Mi corazón ardiendo  
como espigas en la Sierra,  
a la Tierra  
ya es hora de escuchar

Esto pasará, esto pasará  
Tenemos el cielo intacto y la tormenta terminará  
Esto pasará, esto pasará  
Es sólo cuestión de tiempo, la música del silencio renacerá  
Esto pasará, esto, esto pasará

<https://www.youtube.com/watch?v=9RWf8NjnxGY>